

Facing up to endangered apes

A United Nations-sponsored meeting in Paris this week will indicate whether humanity has the wherewithal to save our closest cousins in the animal kingdom from extinction.

As any orang-utan who has looked a human straight in the eyes will know, we have something, well, quite orang-utanish about us. Evidence increasingly points to the great apes — orang-utans, gorillas and chimpanzees — having a range of 'human' attributes, such as culture, emotion and complex social interaction, not to mention being highly intelligent and a marvel to watch. If we allow the great apes to go extinct, we will create a missing link in the understanding of who we are and where we came from.

But all three genera of great apes are climbing up the scale of extinction risk, with most populations somewhere between 'endangered' and 'critically endangered'. Behind these tags lies the stark reality that numbers are now dropping precipitously. Unless we act soon, every species of great ape will be extinct in the wild in our childrens' lifetimes.

Gorillas live in ten African countries, orang-utans on the islands of Borneo and Sumatra in Indonesia and East Malaysia, and chimpanzees in 21 countries in Africa. But by 2030, just 10% of virgin habitat for apes in Africa will still exist, and less than 1% of it for orang-utans.

Destruction of forests is one cause of the decline; hunting for bushmeat or for the live-animal trade is another. And apes have the misfortune to tend to live in war zones, which hampers conservation efforts. Even in remote areas, away from their number-one enemy — us — apes are being wiped out by the Ebola virus (see *Nature* **422**, 551; 2003).

In Paris on 26–28 November, representatives of African and Asian states with ape populations are due to meet with scientists under the

auspices of the United Nations' Great Apes Survival Project (GRASP). They hope to thrash out a conservation strategy for the great apes, endorsed by everyone with an interest in their survival.

There is certainly no shortage of interest. Numerous organizations exist to save them. On paper, great apes are already protected by law in every country they inhabit, national action plans exist, and the international ape trade is banned. But the numbers just keep dwindling.

The meeting's organizers acknowledge that action is needed. Existing conservation efforts are inadequately coordinated and too piecemeal, and projects are set up as funds become available, rather than as part of an overall strategy. The meeting will attempt to expand the ability of GRASP to oversee and implement such a strategy.

Another proposal is to create an International Great Ape Commission — recognized by established zoological bodies — to bring together the countries affected, donors, scientists and non-governmental organizations, to generate publicity for the cause, and to develop common plans and more rigorous systems for evaluating best practice in conservation approaches.

The attention that the meeting will bring to the issue is welcome. The United Nations reckons that a serious effort to lift the immediate extinction threat hanging over great ape populations would cost some US\$25 million. It would be ironic if, just when humans have sequenced our own genome, we allowed a group of species that share almost 99% of it to go extinct. ■

Eastern promise

Untapped scientific potential to the east offers short-term challenges for the European Union, but will strengthen it in the end.

The European Commission's latest report on science, technology and innovation, published on 25 November, makes for a sobering read. It shows not only that the 15 European Union (EU) countries have made little progress towards their avowed goal of increasing their expenditure on research and development (R&D), but that the gap in such spending between the EU and the United States is actually widening.

However, a close reading of the document suggests at least one route forward as Europe struggles to strengthen its research base — the largely untapped human potential of the nations of eastern Europe, most of which are now in the process of joining the EU.

A goal of boosting R&D spending to 3% of the economy by 2010 was publicly proclaimed by Europe's leaders at the Barcelona EU summit in 2002, but now seems to be less realistic than ever. Any prospect of boosting this investment has been choked by Europe's recent economic weakness. In 2001, the most recent year for which a reliable comparison is available, for example, industry and government in the EU 2001 spent €141 billion (US\$165 billion) less on R&D than did the United States — almost twice the gap that existed in 1999.

It is sometimes said that EU enlargement, under which ten eastern and central European nations will join the union next May, will aggravate the discrepancy. But this month's report suggests that in the longer term, the dynamics and favourable demography of the ten

countries might provide a badly needed boost for European science.

Currently, their total R&D expenditures account for just 2% of the EU's €178-billion total annual expenditure. But their growth potential is considerable. Estonia, Hungary, Lithuania, Slovenia and Latvia, for example, are already among the top-ranked economies in the world in terms of the growth of their research activities. And the first results of their participation in the EU's Sixth Framework Programme indicate growing competitiveness in their research capabilities.

The European Commission needs to step in to support the east's budding capabilities, particularly in promising areas such as computer science. In the short term, a change in the rules to allow universities and research laboratories to benefit from so-called 'structural' funds, directed at infrastructure in poor regions, would help.

In the medium term, EU policies should change to accommodate the fact that the main strengths of countries such as the Czech Republic, Poland and Hungary lie not in technology development, where existing policy has its focus, but in basic research. Putting more money into this — through a proposed European Research Council, for example — would help these countries to assert their potential.

As the report makes clear, these countries have great human potential. Their populations are well educated, particularly in science and mathematics. In the long term, Europe must exploit this talent if it is to be globally competitive in science, technology and innovation. ■





Making a splash
GloFish shed light
on loophole in
transgenic rules
p372



On the attack
Promega and Roche
rekindle battle over
enzyme patent
p373



False pretences
University centre takes
benefactor's name
despite faked results
p374



In a spin
Russian lays claim
to invention of
magnetic imaging
p375

State department woos weapons researchers in bid to rebuild Iraq

Geoff Brumfiel, Washington

The US Department of State is launching a fast-track programme to employ former Iraqi weapons scientists in the effort to rebuild the country's shattered infrastructure.

The \$2-million initiative won government approval earlier this month, *Nature* has learned, and state-department officials will be travelling to Iraq within the next six weeks to begin talks with Iraqi scientists. The money will be used to start a series of pilot projects that officials hope will form the basis of larger programmes to redeploy weapons experts and civilian scientists in Iraq.

Ultimately, the state department may spend about \$20 million over the next two years to help restock laboratories and rehire Iraqi researchers who worked on nuclear, chemical and biological agents during Saddam Hussein's regime.

The pilot programme was warmly welcomed by proliferation experts, who had feared that such efforts were getting bogged down amid growing security problems in Iraq and internecine disputes inside the US government.

"I've wanted this kind of thing to happen," says David Albright, a former nuclear-weapons inspector in Iraq who now heads the Institute for Science and International Security in Washington DC. Albright says that the programme has been delayed for months because the Pentagon refused to let state-department officials into Iraq after major hostilities ended on 1 May. In the intervening time, the defence department detained and interrogated many researchers in its hunt for weapons of mass destruction (see *Nature* 423, 371; 2003). "A lot of the scientists were being treated as criminals," Albright says.

State-department officials say that the pilot programme will approach the scientists as allies, and employ them in rebuilding the country. Although many specifics have yet to be worked out, an official at the department says that scientists will work on projects of immediate need such as water desalination, agriculture and rebuilding Iraq's health system.

Two significant challenges will be to



Under the gun: Iraqi academics head into Baghdad University, but the US government is increasingly keen to harness their skills.

identify scientists and engineers who were involved in Iraqi weapons programmes, and to win their cooperation. Unlike in the former Soviet Union, where many weapons researchers were sequestered in secret cities, Iraqi weapons scientists were employed at different times in diffuse locations. Many of them left the government after the 1991 Gulf War to work in industry. The state department is working with the recently established Iraqi Ministry of Science and Technology to identify and approach researchers with relevant knowledge.

Those who have travelled to Iraq in recent months say that scientists there are desperately in need of help. One state-department official visited the science deans from a dozen Iraqi universities in October. After the sanctions that followed the first Gulf War, many scientists lost contact with the broader international community, and the younger generation of researchers has been poorly trained, the official says: "They're going to need a lot of retraining and continuing education."

George Atkinson, the science adviser to

secretary of state Colin Powell, is meanwhile preparing a separate proposal aimed primarily at civilian scientists. Known as the Science, Technology and Engineering Mentorship Initiative for Iraq (STEM II), this plan would provide almost \$16 million to put about 3,000 scientists in the country back to work. In addition, US mentors would be enlisted to bring the Iraqis up to speed with trends in their respective disciplines.

Unlike the current programme, which deals only with former weapons scientists, STEM II will reach out to the entire scientific community of Iraq. "STEM II is about giving a productive scientific future to people not directly involved in weapons of mass destruction," says a state-department official who helped to draft the proposal.

Experts agree that in order for both programmes to be effective, they will need to overcome several obstacles, the foremost of which is security. The continued attacks against the US-led coalition and Iraqis who support it have created a climate of fear, according to Michael Roston, a senior analyst at the Russian American Nuclear Security Advisory Council in Washington DC. "I think it will be a real uphill battle to make Iraqi scientists confident," he says.

The security situation will also make it dangerous for US researchers and state-department officials trying to meet with their Iraqi counterparts, adds Carol Kessler, director of the Center for Global Security at the Pacific Northwest National Laboratory in Richland, Washington.

But even if they prove difficult to implement, programmes such as these are long overdue, says Albright. In the months since major fighting ended in Iraq, many scientists have gone into hiding or fled the country. "The United States has lost a lot more than just time," he says. ■

RABIH MOGHRABI/AFP/GETTY IMAGES

Germany puts faith in big guns as science ministry feels pinch

Quirin Schiermeier, Munich

Germany's research spending will be trimmed next year by about €100 million (US\$120 million), dashing hopes that science can escape the impact of the nation's growing fiscal difficulties.

In a plan scheduled to be approved by parliament on 28 November, Germany's main research agency, the DFG, and three large research organizations, including the Max Planck Society (MPS), will each get a budget increase of 3% in 2004 — lifting this year's budget freeze.



Whistle-blowing: Munich students protest against cuts in the state education budget.

The DFG and MPS will receive about €1.3 billion and €960 million, respectively, from the research budget. But the overall budget of the education and research ministry, which also provides grants in areas such as biotechnology and genetics, will be cut by €100 million — about 4%.

The cuts at the research ministry will “drastically increase” the demand for DFG grants, says Ernst-Ludwig Winnacker, the agency's president.

Dark clouds are also gathering over universities in Bavaria, Germany's richest state, where science has enjoyed strong support. On 20 November, tens of thousands of students in Munich and other cities protested against state plans to cut university budgets by 10% next year.

“In the long term this would force us to close several institutes,” says Bernd Huber, the rector of Ludwig-Maximilians University in Munich, which has to save €30 million next year. “But in the short term, there's nothing we can do but cease to renew young scientists' contracts.” ■

GloFish casts light on murky policing of transgenic animals

Jonathan Knight, San Francisco

A fluorescent fish that will be on sale in pet stores has exposed a gaping hole in the United States' ability to regulate the sale of genetically engineered animals.

Despite growing concern over the potential risks of releasing such animals into the environment, the transgenic GloFish looks set to go on sale on 5 January without any federal regulatory approval.

Watchdog groups and some biologists warn that this sets a worrying precedent. Some aquarium fish are invariably released into the wild, they say, and even if the GloFish is safe, the floodgates will be opened to riskier transgenic pets. “This isn't just one little fish, it's a parade,” says Margaret Mellon of the Union of Concerned Scientists in Washington DC.

But the company preparing to sell the fish, Yorktown Technologies of Austin, Texas, says that it has consulted numerous scientists to ensure that the GloFish does not pose a threat. “I would encourage any company following us to do the same,” says chief executive Alan Blake.

No federal agency has claimed jurisdiction for such animals, because the US government has no specific regulations for the use of transgenic organisms, having decided in 1986 simply to adapt existing rules. The resulting framework divides jurisdiction among three agencies — the Environmental Protection Agency (EPA), the Food and Drug Administration (FDA) and the Department of Agriculture — depending on the gene inserted.

Years later, when Aqua Bounty Farms of Waltham, Massachusetts, proposed inserting a growth-hormone gene into salmon, the rules were stretched to cover the fish by classifying any added protein that changes an animal's function as a drug to be regulated by the FDA. The agency is now considering Aqua Bounty's application to farm the transgenic salmon.

But no rule-stretching has yet been done to cover pet fish. Blake says he contacted the FDA, which told him that the GloFish does not fall under FDA jurisdiction. He reports similar responses from the EPA, the Department of Agriculture and the Fish and Wildlife Service.

Scientists who have seen data on the GloFish say it does not seem to pose an environmental threat. It was developed by Zhiyuan Gong at the National University of Singapore, who plans to link the colour genes to genetic elements that respond to pollu-

tants such as heavy metals, to monitor water quality. When Gong tested the fish's reproductive success in mixed populations, the transgenic fish had fewer offspring than wild ones.

That suggests they are safe for release, says William Muir, an aquaculture expert at Purdue University in West Lafayette, Indiana, one of several scientists providing Blake with unpaid advice. “You might think that something glowing would have an advantage, but it has the reverse effect,” he says.

But Muir says that such unexpected findings only underscore the need for careful regulation of transgenic animals, because it is impossible to predict a transgene's effects. Muir was the first to describe the ‘Trojan gene’ effect, in which an introduced gene spreads rapidly in a population while at the same time reducing overall survival rates (see *Nature* **406**, 10–12; 2000). “There should definitely be a regulatory process,” he says.

The only government review so far is taking place in California, where in May the state's Department of Fish and Game banned the possession, sale and transport of genetically engineered fish. Yorktown has applied for an exception for the GloFish, and the department is preparing to present a recommendation at the next Fish and Game Commission meeting on 4 December.

But federal involvement is essential, says Peter Jenkins of the Center for Food Safety in Washington DC, one of five groups petitioning the FDA to regulate the fish. “Glowing reviews by a few scientists are not enough,” he says. “There should be a formal process.”

Anne Kapuscinski, a fisheries expert at the University of Minnesota, Twin Cities, says the GloFish is an opportunity for the FDA to establish its authority in the transgenic pet trade. Approval would be fairly straightforward, she thinks. “It behoves the government to act as proactively as possible,” she says. “People are losing trust in the process.” ■



Glowing review: watchdogs want tighter rules for transgenic pets.

Promega changes tack in battle over patent

Jonathan Knight, San Francisco

A company that sells an enzyme used in the polymerase chain reaction (PCR) is suing the corporation that holds a patent on it, alleging that researchers who use the enzyme almost every day are being overcharged.

Promega of Madison, Wisconsin, has filed a lawsuit against the Swiss pharmaceutical giant Hoffmann-La Roche, which holds a patent on the *Taq* enzyme. In the suit, which was filed on 13 November at a federal district court in Virginia, Promega alleges that government-funded researchers, who use the enzyme in PCR to amplify genetic material, are paying Roche too much.

Promega is claiming \$1 billion under a law that allows you to file a suit against anyone you think is swindling the US federal government. If successful, the company stands to collect up to 30% of any award. Roche says that the allegations are "without merit".

The case is being pursued under the 1863 False Claims Act, which protects the govern-



PCR is at the heart of gene-sequencing — and patent litigation.

ment against fraud by allowing citizens to sue on the government's behalf. A 1986 amendment allowed the litigant to claim treble damages and raised the maximum amount that the litigant may pocket from 25% to 30% of the total. Since then, the

number of suits under the act has soared, and the US Treasury has recovered more than \$10 billion.

"It certainly is an innovative tactic," says Greg Aharonian, a technology patent consultant based in San Francisco. Of half-a-dozen intellectual-property lawyers contacted by *Nature*, none had heard of the act being invoked in a patent dispute.

Roche and Promega have been wrangling over *Taq* since 1992, when Roche sued Promega for violating its patent on the enzyme, which is isolated from the bacterium *Thermus aquaticus* that lives in hot springs. Roche obtained the patent in 1991 from Cetus, a now-defunct biotechnology company in Emeryville, California. Promega responded to the suit by claiming that the patent had been fraudulently obtained and should be invalidated.

In 1999, the US district court in San Francisco finally agreed with Promega. In his decision, Judge Vaughn Walker ruled that the original patent application from Cetus contained eight instances of misrepresentation that were intended to deceive the patent office, rendering the *Taq* patent unenforceable.

Roche quickly appealed. In March 2003, the US Court of Appeals for the Federal Circuit in Washington DC, overruled Walker on six of the eight cases. The appeals court then sent the case back to Walker to decide whether the remaining two instances were sufficient to justify voiding the patent. A decision is expected early next year.

One reason that the whistleblower act hardly ever comes up in patent cases is that the government rarely spends a lot of money on patented products, Aharonian says. But over the years, government agencies have spent a small fortune on *Taq*, including royalty payments to Roche.

Promega reached the sum of \$1 billion by claiming triple damages, as allowed by the law, plus a \$10,000 charge for each time a *Taq* purchase generated royalties for Roche. "This looks like a chance to go for big bucks," says Jerry Selinger, a Dallas patent lawyer and board member of the American Intellectual Property Law Association.

But Selinger and others don't give the suit much chance of success. For one thing, the Department of Justice, after reviewing the lawsuit's claims, declined to participate in prosecuting it. And although the future of Roche's *Taq* patent is uncertain in the United States, the European Patent Office issued a final ruling upholding it on 30 October.

US draws up plans to tackle autism

Emily Singer, Washington

The National Institutes of Health (NIH) is preparing a research plan for autism — and hopes to cut the condition's prevalence in the United States by a quarter by 2013.

The NIH autism 'roadmap', revealed on 19 November at a planning meeting of researchers and autism activists in Washington DC, will aim to identify the genetic, environmental and neurological factors behind the disorder. It represents the biomedical research agency's first concerted push to tackle the condition, which afflicts up to 1 in 200 people in the United States.

Requested by the Congress last year, the roadmap was drawn up by the Interagency Autism Coordinating Committee, a group of scientists, autism advocates and public-health experts. It may form the basis of future budget requests from the National Institute of Mental Health (NIMH), which will lead its implementation.

By setting specific benchmarks for treating the disorder, the plan is something of a departure for the NIH, reflecting the results-oriented approach advocated by its director, Elias Zerhouni (see *Nature* 425, 438; 2003). The main scientific aspects of the roadmap include building a database of genes potentially linked to autism, and a series of clinical trials to determine the effectiveness of various early treatments for the condition. But the plan also sets what it calls a "high-risk" goal of reducing incidence of the disorder by 25% within ten years.

Researchers haven't tackled these problems before, partly because so little is understood about the mechanisms behind autism. There is currently no biological diagnosis for the disorder. "We have no genes, no circuits, no workable animal models, so we don't have the tools to develop new treatments," says NIMH director Thomas Insel. "It's a striking contrast to where we are with the rest of medicine. We are where we were eight years ago with Alzheimer's disease or 20 years ago with Huntington's."

At this point, NIH officials don't know how much money will be available to support the initiative. But some of the work will be done at eight autism research centres, established last year by the NIH, with a total budget of \$65 million over five years.

The NIH and the National Alliance for Autism Research, an advocacy group based in Princeton, New Jersey, are already supporting a \$4.5-million project to coordinate the search for autism-related genes. Now advocacy groups hope that the roadmap will form the basis of a larger, publicly supported investigation of the condition.

"It's gratifying to see the government addressing it at such a high level and with such high priority," says Lee Grossman, chairman of the Autism Society of America, an advocacy group based in Bethesda, Maryland. "It's what we've been asking for for decades."

University buildings named on shaky ground

Rex Dalton, San Diego

Before his death last year, the legendary archaeologist Frank Hibben made a \$3.5-million donation to the University of New Mexico to fund an extension to its renowned anthropology museum.

But as the Hibben Center for Archeological Research nears completion in Albuquerque, it is confronting a very different aspect of Hibben's legacy: his apparent faking of research over many decades.

Some faculty members at the university are using Hibben's transgressions — which have been widely publicized for years — to teach students how not to conduct themselves. Researchers in the field even refer to data contortions as 'hibbenisms'. "He thought it didn't hurt to make the evidence a little better," says Vance Haynes, an archaeologist at the University of Arizona in Tucson.

American universities constantly seek major benefactors for building projects, with staff at development offices working to bring in donations. Most contributors have had stellar careers. But sometimes donations are made by individuals or firms who are seeking legitimacy after miring themselves in controversy.

For instance, the first laboratory building at the new Mission Bay campus of the University of California, San Francisco (UCSF) is named Genentech Hall. What few of its students may realize is that the building was paid for by a \$200-million lawsuit settlement after San Francisco-based Genentech allegedly stole valuable human growth-hormone DNA from UCSF. Genentech made the payment to UCSF in 1999 without admitting any wrongdoing (see *Nature* **402**, 335; 1999).

At the University of Arizona, meanwhile, an agricultural-science building was named in 1992 to honour land baron Kemper Marley — the reputed mastermind of the 1976 bombing murder of Don Bolles, an investigative reporter for the *Arizona Republic*. When Marley died, his family made a \$6-million contribution to The Marley Building, a name that still infuriates some people. "I refused to go into the building," says Don Carson, a former journalism professor at Arizona.

At Seton Hall University in South Orange, New Jersey, the name of businessman Robert Brennan, who had been convicted of fraud, was physically chiselled off a recreation centre late last year after the university decreed a revised policy for the naming of its buildings.

"These cases are rare," says Sheldon Steinbach, an attorney for the American Council on Education (ACE), which represents 1,800 academic institutions. But after a recent spate of corporate malfeasance in the United



The University of New Mexico's archaeology centre honours Frank Hibben (below), who faked results.



States, he says, "universities are likely to look a little more askance at large donations from sketchy sources".

Neither ACE nor the Association of American Universities, which represents top research universities in the United States and Canada, has a clear policy on building donations, because each case is so different.

The reason for Hibben's fame was his 1937 claim to have discovered undisturbed sedimentary layers showing that people lived in Sandia Cave, New Mexico, about 25,000 years ago. This is by far the earliest date claimed for an indigenous American culture.

Hibben co-founded the University of New Mexico's Maxwell Museum of Anthropology in 1932 and led it for three decades, during a golden age of discoveries in the

heritage-rich southwestern United States. A charismatic professor at the university, he was a mentor to many of today's leading US archaeologists, and is widely credited with popularizing modern archaeology.

Scientific publications over the past 25 years have questioned or disproved several of his most noted discoveries. His Sandia Cave claim, for example, was discredited by Haynes and others, who found the cave's rock formations to be quite different from those described by Hibben. But he was never publicly accused of scientific misconduct.

In 1946, Hibben described an expedition to Chinitna Bay on the west side of Cook Inlet in Alaska. He claimed that a geological site in the bay has lithic points that match those of the Folsom people, who are known to have lived in New Mexico 10,000 years ago.

But in 1978, when archaeologist James Dixon and others returned to the isolated location, they found Hibben's claim to be false. "It is clear that the site never existed," says Dixon, now at the University of Colorado in Boulder. In classes at the University of New Mexico, archaeologist Bruce Huckell provides articles on these interpretations to students. "Some think Hibben's work is all faked," says Huckell. "Others think we don't have enough information to know."

At UCSF, medical anthropologist Judith Barker says that she uses such cases to teach responsible research conduct. She says that she has not used the Genentech case as an example, but may yet do so. "A lot of young scholars don't understand that they don't own the research they conduct at a public university," she says. "It can be an unpleasant lesson to learn. They could learn from the Genentech case."



Russian claims first in magnetic imaging

Bryon MacWilliams, Moscow

At the height of the cold war, Vladislav Ivanov was a young lieutenant in the Red Army charged with trying to use nuclear magnetic resonance in water to help with aircraft navigation.

The days were long and lonely at his radar station, Ivanov recalls, and he had plenty of time to think. Time enough, according to documents filed in 1960 with the USSR State Committee for Inventions and Discovery at St Petersburg, to think of the idea of using magnetic resonance to study the inside of the human body.

"I figured that, because a person is made up primarily of water, the same method could be used in research on living organisms," Ivanov says. "The water inside a person could be used to give a signal showing what exists, or is located, inside."

This concept — magnetic resonance imaging (MRI) — was recognized in October when Paul Lauterbur of the University of Illinois, Urbana-Champaign, and Peter Mansfield of the University of Nottingham, UK, shared the 2003 Nobel Prize in Physiology or Medicine. Ivanov lays no claim to a share in the prize: his ideas stagnated after his seniors rejected them as unviable.

Still, he chuckles at the public protests of Raymond Damadian, a New York-based physician who claims to have thought of the idea first (see *Nature* 425, 648; 2003).



Vladislav Ivanov's ideas for magnetic resonance imaging were hampered by bureaucracy in St Petersburg.



Damadian, whose work was published in 1971, believes that he should share the prize with the physicists. "If I had never been born, there would never be MRI today," Damadian said after the prize was announced.

"I know Damadian's work, but at this point there's nothing that can be done," Ivanov said last week in his office at the St Petersburg State Institute of Fine Mechanics and Optics. "Besides, there are always mistakes when you have major advances in science. You can't keep an idea in one place. They have their own momentum," he adds.

In Ivanov's case, the momentum was quashed by the state committee in St Petersburg, whose records show that Ivanov submitted an application, "Method of examination of the internal structure of material bodies", in 1960.

The application defined Ivanov's principles in detail, and included a diagram of a device identified as MRI equipment. The document was assigned the number 659411/26, and was forwarded for review to scientific institutes in St Petersburg, then known as Leningrad.

"The committee members thought about it, and couldn't figure it out themselves, so they sent it to a critic, who came out sarcastically against it, then sent it to his boss, who signed off on it," Ivanov recalls. "They thought I was mistaken, but I wasn't."

His options for appeal were limited: Red Army officers could not publish abroad. "After that they even started to spy on me, my commander and everyone else to whom I turned for help. They decided I wasn't doing my job, that I had a lot of free time."

Years later, when the work of Lauterbur and Mansfield came to the world's attention, the committee revisited Ivanov's application and gave its assent. But by then, Ivanov says, Soviet science had fallen too far behind.

Today, all he has to show for his ideas is an MRI machine, based on his own ideas, on which he trains students in his lab. "On one hand, the Cold War is guilty. But it was never so cold that it was impossible to do at least something." ■

US on sidelines as China prepares for satellite launch

Tony Reichhardt, Washington

The first of two Chinese Double Star satellites is set to launch next month, marking the most ambitious scientific mission yet for the nation's space programme.

But while European researchers are vigorously engaged in the mission, which will explore the interaction of the solar wind with Earth's magnetosphere, their American colleagues are staying on the sidelines — highlighting a rift between the world's largest space power and its fastest-growing rival.

The first 660-kilogram Double Star satellite is scheduled to reach orbit on 28 December, and its twin will blast off next June. Data on high-energy particles and magnetic fields returned by the spacecraft will supplement information from four European/US Cluster satellites that are already in orbit.

The European Space Agency (ESA) has spent €8 million (US\$9.5 million) on half of the 16 scientific instruments on board the two satellites (see *Nature* 412, 106; 2001).

Some 75 US scientists received NASA funding to participate in the joint Cluster mission. But when China asked ESA to participate in Double Star in 1999, it did not extend the invitation to NASA. As a result, although seven of the eight European instruments on board are duplicates of Cluster instruments, only a dozen or so US scientists will participate as informal members of ESA teams.

In some cases, Cluster instrument components that had been built in the United States had to be manufactured in Europe for Double Star, to get around US regulations restricting the export of technology with possible military application, such as radiation-hardened parts (see *Nature* 416, 572; 2002).

Solar-terrestrial space research has traditionally been a strong area of cooperation between NASA and ESA. NASA's upcoming ventures in the field, such as the Solar Terrestrial Relations Observatory (STEREO) and the Magnetospheric Multiscale Mission, will still involve international partners, but not as many as before. And the export regulations — officially known as ITAR (International Traffic in Arms Regulations) — have caused further complications.

Mario Acuña, a space scientist at NASA's Goddard Space Flight Center in Greenbelt, Maryland, and an investigator on STEREO, says: "We made it so difficult for Europeans to collaborate with us that we have become a pain that nobody wants to deal with... Every presentation has to be cleared, and the non-US scientists have to have separate meetings. It's an insult to them." ■

UK primate centre gets the nod but faces cash shortfall

London A planned primate-research facility at the University of Cambridge, UK, which has been the subject of a long-running row between the university and animal-rights groups, has been given the go-ahead by the British government.

A proposal to build the Centre for Behavioural Neuroscience, which would use primates for research into Alzheimer's and Parkinson's diseases, was rejected by South Cambridgeshire council last year after police warned that the research could attract violent protests. The deputy prime minister, John Prescott, overruled that decision on 21 November, stating that the centre's planned research is in the national interest.

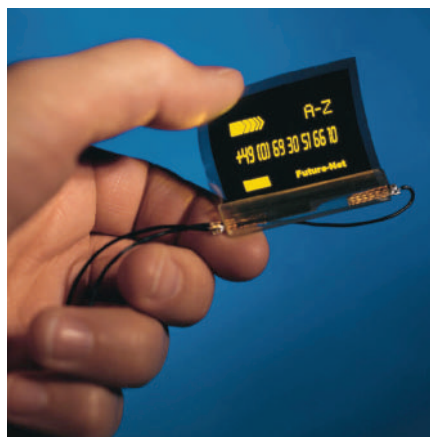
But the centre's future is far from certain. The delay, combined with the need for additional security at the site, has pushed up costs by some £8 million (US\$14 million) to £32 million. The original £24 million pledged by the government and the Wellcome Trust, an independent research-funding charity, remains secure, and the university says it is in discussion with potential funders regarding the shortfall.

Bendy screens and wobbly world win European award

Paris This year's Descartes Prize — a €1-million (US\$1.2-million) purse that rewards successful European collaborations, has been awarded to two teams of researchers.

Richard Friend, a physicist at the University of Cambridge, UK, together with researchers and companies from Belgium, Sweden, the Netherlands and Germany, won €700,000 for discovering that light can be emitted by semiconducting polymers (see *Nature* 397, 121–128; 1999), and for their work to turn these plastics into flat displays for use in roll-up computer and television screens. They were first used last year, in the battery-level display on the Philips Spectra electric shaver.

The other €300,000 went to a consortium



On a roll: the flexible screen that won its inventors a share of the Descartes Prize.

led by Veronique Dehant of the Royal Observatory of Belgium, with researchers from Austria, the Czech Republic, France, Germany, Poland, Russia, Spain and the Ukraine. They developed a model to predict a wobble in Earth's axis, caused by the pull of the Sun and Moon, helping to improve the resolution of global satellite-navigation systems tenfold.

► www.cordis.lu/descartes

Particle-physics funder backs linear collider

London Britain's particle-physics funding body has thrown its weight behind plans for an international linear collider, a multibillion-pound project to study particles produced by high-energy collisions between electrons and positrons.

The endorsement comes in the Particle Physics and Astronomy Research Council's five-year plan, which is due to be released on 25 November. The council will spend £11 million (US\$19 million) on accelerator research and development between now and 2006 to ensure that British scientists are involved in the collider, although it does not expect Britain to host the facility. It also intends to invest in plans for an international neutrino research facility, a project that it says could be hosted at the

Rutherford Appleton Laboratory near Didcot in Oxfordshire.

The collider project has been gaining momentum in recent weeks, although governments are not expected to make a final decision on funding until 2007. On 10 November it was ranked as a mid-term priority by the US Department of Energy (see *Nature* 426, 108; 2003). The following week, a panel to oversee designs for the collider was established at a meeting of the International Linear Collider Steering Committee in Paris.

Congress endorses big spending on small science

Washington The US Congress has passed a bill recommending that almost \$4 billion should be spent on nanotechnology over the next four years, with a National Nanotechnology Coordination Office overseeing research.

The bill also calls for a federal advisory panel to look into the societal and ethical issues surrounding nanotechnology, which has attracted some concern over its possible effects on health and the environment (see *Nature* 424, 246–248; 2003).

The bill, which was unanimously passed by both houses, continues support for the government's National Nanotechnology Initiative, created in 2000. The initiative was originally set to run for five years, but the bill will extend investment until 2007.

"We think these are positive steps," says Mark Modzelewski of the NanoBusiness Alliance, a group based in New York City that represents more than 250 US research firms interested in nanotechnology.

Irish scientists smiling at research cash boost

London Ireland's basic research has been boosted by a 62% increase in next year's budget for Science Foundation Ireland, the country's main science agency. The government has also lifted a freeze on funding for the Higher Education Authority, which supports university science.

The changes are good news for a country that has experienced both drought and plenty in its science funding. After stagnating for years, Ireland's spending on basic research increased by an order of magnitude in 2000 (see *Nature* 404, 215; 2000). But this was set back by a freeze on Higher Education Authority funding which was announced in November 2000.

Science Foundation Ireland's budget will increase by €44 million (US\$52 million) to €113.7 million for 2004. The increase was welcomed by the foundation's director, William Harris, who says it is crucial to sustaining Ireland's position as a "globally competitive, knowledge-driven economy".

Polar eclipse may cast light on solar size

London Shadow-chasers, who travel the world in pursuit of eclipses, had a long way to go last week: the total solar eclipse (right) that occurred on 23 November was visible only from the Antarctic.

Several scientists made the trek south to watch the event, including astronomer John Parkinson from Sheffield Hallam University, UK. He monitored the duration of the eclipse as part of an ongoing



effort to find out whether the Sun is measurably shrinking.

"It was absolutely brilliant," he told *Nature*, from an icebreaker parked in the path of totality. "The corona seemed particularly active," he said, adding that this may

have been linked with some particularly large solar flares seen a month earlier. "The Sun spins once every 26 days, so we saw it about one full spin after these big blobs were ejected."

Eat your veg

With fish farming on the rise, researchers are seeking ways to make aquaculture more sustainable. One solution may mean turning carnivorous fish into vegetarians. Kendall Powell gets a taste of the future.

Global aquaculture is on the rise, growing more than 5% per year over the past decade. That might sound like good news for the world's food supply, but there's a hidden cost behind some of the farmed fish on supermarket shelves. Many, including the popular salmon, trout and cod, are fed on wild fish. Lots of wild fish. Today, about 11 million tonnes of fish — 12% of the total haul from seas and rivers — are caught each year just to feed farmed fish^{1,2}. It takes 2 to 5 kg of wild fish just to produce 1 kg of a farmed fish such as salmon³.

Over the past few decades, researchers have begun to think that one way to make aquaculture more sustainable is to change the diets of some of our farmed fish — to turn carnivores into vegetarians³. It's a solution that carries its own challenges, but in the face of declining wild stocks and a booming aquaculture industry, many fish farmers and conservationists agree that if we are to continue farming carnivorous fish, this is the way to go.

The world's most-farmed fish, carp and tilapia, are predominantly vegetarian, and so can easily be farmed on a diet of plants. In Asia, where currently about 90% of the

world's aquaculture takes place, these fish are big business. But in the western world, where carp is often considered to be bony and tilapia is a relatively new and unfamiliar product, neither has so far managed to out-compete meaty salmon steaks on restaurant menus.

Lagging just behind the production of farmed tilapia are salmon and trout — both of which are natural carnivores. Although the number of these fish being farmed is one-tenth that of farmed carp, it is a growth area, and farming of other carnivorous fish, such as cod, halibut, sea-bass, sea bream and tuna, is also increasing. Only 600 tonnes of farmed Atlantic cod were produced worldwide in 2001, but Norway has now started commercial production, and Hans Abrahamsen, managing director of aquafeed producer Skretting in Stavanger, Norway, estimates that the country will produce up to 30,000 tonnes per year by 2008. Altogether, the production of carnivorous fish is expected to double by 2010 (ref. 2).

This will soon pose a huge problem. Farmed fish are fed on a diet that leans heavily on fish oil and fishmeal — a protein-rich powder of ground-up, cheap fish such as

sardines, anchovies and eels — as a source of vital proteins and nutrients. A simple calculation shows that the current haul of fish oil and fishmeal will soon be outstripped by the needs of global aquaculture. If the number of fish farmed continues to grow at its current rate, and if the supply of oil and meal stays the same — as it has for the past decade — then demand will outstrip supply of oil by 2010 (ref. 4). If those projections are extended, fishmeal looks set to face the same problem by 2050.

Feeding frenzy

In response, feed companies are seeking different sources of protein and, in particular, oil. The feed fisheries have turned to larger fish, such as mackerel, herring, blue whiting and Norway pout, causing further pressures on natural stocks. Some companies are also going after krill — tiny crustaceans found predominantly in Antarctica. "That's going to the base of the food chain, which has all kinds of ramifications for the oceans' web of life," says Bill Mott, director of the information centre SeaWeb Aquaculture Clearinghouse in Providence, Rhode Island.

There are also economic reasons to switch

R. MORSCH/CORBIS



What's on the menu? Many fish, such as this bass, are carnivorous. But farmers hope to turn fish-eaters such as salmon into vegetarians.

diets away from fishmeal and oil. Feed makes up three-quarters of fish farmers' total running costs, and the price of oil will only go up as demand rises. "For fish-feed producers today, it's not a question, we have to use alternatives," says Jørgen Holm, a fish nutritionist at the feed company BioMar in Brande, Denmark.

In theory, this shouldn't be difficult. Trond Storebakken, director of the Aquaculture Protein Centre in Ås, near Oslo, Norway — an institute devoted to this problem — says that moving a carnivorous fish onto a vegetarian diet is similar to someone deciding to become a vegan. "A person eating only boiled soya beans and raw cabbage would get bone demineralization and goitre in a few months," he says. "But if you know the requirements then you can eat the right things."

Hard to swallow

Today, the options for vegetarian fish diets include soya beans, corn, rapeseed, sunflower seeds, flaxseeds and wheat gluten. Some farmed salmon already get up to half of their protein or oils from such ingredients, without their health or growth rates suffering⁵. But attempts to push diets past the half-vegetarian point tend to cause serious digestion problems — fish suffer from irritated lower intestines and a depressed immune system, and lose their ability to absorb key minerals such as zinc and iron. So far, the only way to solve these problems costs a lot of money. "I could give you a simple prescription for a 100% vegetarian diet for carnivorous fish that works, but it would be extremely expensive," says Storebakken.

If you wanted salmon to survive on nothing but soya beans, for example, you would have to supplement their diet with missing essential amino acids, such as methionine. More importantly, you would have to find a way to break down phytic acid, the major storage molecule for phosphorus in seeds. Fish don't make phytase, the enzyme that breaks down the acid and allows grazing animals to extract the nutrients from plants. The enzyme can be added to fish feed⁶, but a special heat-stable version — isolated from a fungus gene and produced in bacteria — is required to work efficiently throughout feed processing. It remains to be seen whether this method will be cost-effective.

Plants also contain 'antinutrients' that can be toxic to fish in high concentrations, or that can decrease the fish's ability to absorb minerals from their food. Antinutrients vary from plant to plant and their effects differ for each species of fish, so the diet needs to be carefully tailored for the fish in question. With about 200 species of farmed fish, such



Weight loss: it can take up to 5 kg of wild fish to produce 1 kg of farmed salmon.



requirements are a financial headache.

Even if you can get fish to eat an all-vegetarian diet, there is the question of whether vegetarian fish taste the same as a carnivorous ones. For better or for worse, using only vegetarian oils in the diet results in milder-flavoured, less-oily fish fillets with a different texture. But experts disagree on whether consumers will notice. Abrahamsen says that in taste tests run by Skretting, professional food tasters couldn't tell the difference between salmon raised on vegetarian diets and those raised on traditional carnivorous fare.

Grand designs

More importantly, vegetarian diets tend to reduce the amount of omega-3 fatty acids in fish flesh — the compounds that reduce the risk of heart disease in people who eat them. But many feed researchers see this as a challenge rather than a problem. "Just by simple additions to the diet, we can produce a 'designer fish'," says Ewen McLean, director of the Virginia Tech Aquaculture Center in Blacksburg. In pilot studies, McLean has added the enzyme lipase, which breaks down fats, to fish diets and has seen an increase in the omega-3 fatty-acid content of the fish⁷.

Another way to rectify the fat content is to give a fish a vegetarian diet for most of its life, and then follow it with a more 'natural', fishier diet just before going to market. Paul Brown, an aquaculturist at Purdue University in West Lafayette, Indiana, says that just three weeks on such a finishing diet could restore traditional flavours to salmon raised on a vegetarian diet

for 18–24 months. Other experts say that a natural diet might be required for two to three months to ensure high levels of omega-3 fatty acids. Even so, that would cut the amount of fish oil used over a fish's life by about 85%.

Many researchers believe that the final hurdles to making vegetarian feed realistic and affordable will be cleared in the next 5–10 years, allowing fish diets to be close to 100% vegetarian. Three major fish-feed companies — BioMar, Skretting and Ewos, based in Bergen, Norway — have pledged to replace at least 50% of the fishmeal in their feeds with alternative protein sources by 2010, in addition to finding alternative oil sources.

There are other advantages in store for these farms if they adopt a more vegetarian feed. BioMar, for example, has found that fish fed a diet containing a greater proportion of vegetarian oils tend to excrete less phosphorus and nitrogen in their waste. These are two of the main culprits that cause water pollution in fish farms, as they fertilize algal blooms that can take over ecosystems, using up oxygen and suffocating fish. Researchers working on custom-designing their feed think that these vegetarian diets can be tweaked to optimize this effect, reducing pollution even further.

Large-scale fish farming will still face other obstacles, including problems of escaped fish mating with natural stocks in the wild, and the outbreak of disease in the close quarters of farmed stocks. But if researchers can at least wean carnivorous fish off a diet of wild fish, they will be one step closer to a sustainable industry.

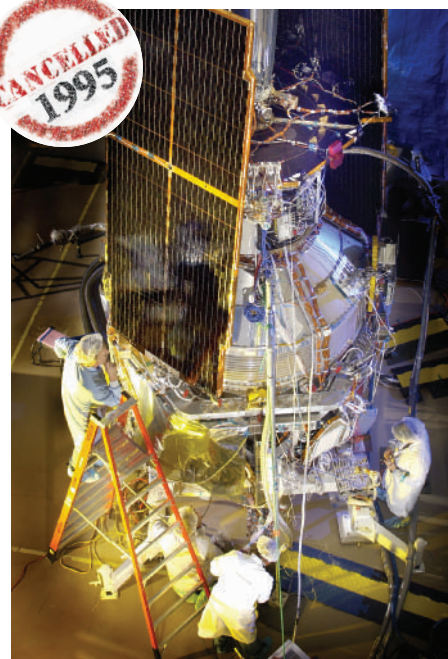
If they succeed, pressures on wild fish stocks won't skyrocket as aquaculture booms. At the same time, the new diet will hopefully reduce pollution, the cost of fish — and even world hunger, as at least some of the small fry currently fed to farmed fish could instead be used to feed people. If aquaculture as a whole is to succeed, it needs at the very least to be sustainable — not only for the benefit of wild stocks, but ultimately for the industry and consumers too.

Kendall Powell is a science writer in Broomfield, Colorado.

1. FAO Fisheries Department *The State of World Fisheries and Aquaculture 2002* (FAO, Rome, 2002); www.fao.org/sofi/sofia/index_en.htm
2. Weber, M. *What Price Farmed Fish* (SeaWeb Aquaculture Clearinghouse, Providence, RI, 2003); www.seaweb.org/resources/sac
3. Naylor, R. L. *et al. Nature* **405**, 1017–1024 (2000).
4. International Fishmeal and Fish Oil Organisation *Sustainability of Fishmeal and Fish Oil Supply* presented at the Scottish-Norwegian Conference on Sustainable Futures for Marine Fish Farming, June 2001; www.iffco.org.uk/tech/sterling.htm
5. Refstie, S., Storebakken, T., Bæverfjord, G. & Roem, A. J. *Aquaculture* **193**, 91–106 (2001).
6. McLean, E., Craig, S. R., Goddard, S. J., Al-Oufi, H. S. & Wille, K. *Ribavirto* **60**, 15–28 (2002).
7. Samuelsen, T., Isaksen, M. & McLean, E. *Aquaculture* **194**, 161–171 (2001).

Unstoppable force

After 40 years in development, and some \$650 million of NASA funds, Gravity Probe B is almost ready to launch. Would Einstein, whose theories it is about to test, have approved? Tony Reichhardt reports.



Charmed life: three times during the twisting saga of Gravity Probe B's construction, NASA has tried unsuccessfully to pull the plug on the project.

At first glance, the saga of Gravity Probe B looks like a success story. Plucky scientists and engineers have spent four decades — most of their working lives — fighting technical and political battles, beating huge odds to reach the launch pad at long last. It is the longest-running development project in NASA's history. Some observers think it's a colossal waste of money, and the space agency has tried to cancel the mission at least three times. But the US Congress has always restored its funding, and despite yet another setback on the launch pad this month, the probe is now set to make a triumphant launch in the first half of next year.

Even the critics have to acknowledge Gravity Probe B's technical brilliance and the sheer gutsiness of its engineering. It will use the most perfect gyroscopes ever built to make one of the most sensitive measurements in the history of physics. As it passes from pole to pole in its 640-kilometre-high orbit, it will test two key predictions of Einstein's general theory of relativity: that a massive body such as Earth warps the fabric of space-time, and that it drags space-time around with it as it rotates. Neither prediction has been proved experimentally, at least not to Gravity Probe B's level of precision. And if by chance the experiment throws up a surprise, it could have profound consequences for physics.

So why, just a few months ago, was NASA's head of space science so frustrated that he threatened, not for the first time, to kill the project?

Heavenly vision

First, some history. Gravity Probe B dates from 1960, when physicist Leonard Schiff and colleagues at Stanford University proposed an experiment that would exploit advances in gyroscope technology to look for evidence of 'frame-dragging'. They pictured an orbiting telescope holding its gaze on a target star while four almost perfect gyroscopes spin alongside the instrument, their axes all oriented in the same direction. If frame-dragging were to tug at the gyroscopes as Einstein predicted, their axes would shift ever so slightly over the course of a year, allowing physicists to measure the phenomenon with a precision of 1% or better.

By 1964, NASA was providing low-level funding for the project. Perfect quartz spheres had to be fashioned and electrically suspended at the heart of each gyroscope so that they could spin freely in a vacuum. The design would require an ultra-cold container filled with superfluid helium, and ultra-gentle thrusters to manoeuvre the spacecraft — everything on Gravity Probe B had to be 'ultra'. Within 20 years, the technology had pro-

gressed enough for NASA to fund a test flight, bringing the project's price tag to \$130 million, not counting the shuttle's launch costs.

By the late 1980s, however, NASA was reeling from the Challenger accident, money was tight, and the agency was increasingly relying on 'decadal reviews' by the National Academy of Sciences to set funding priorities. An expensive gravitational physics experiment was not high on many people's agenda. Even today, would Gravity Probe B survive a peer-reviewed competition with other space experiments in its price bracket? "Probably not," says Eugene Parker, a University of Chicago astrophysicist who kept faith with the project when chairing a review of it in 1991.

NASA cancelled Gravity Probe B in 1989. It cancelled it again in 1993, and again in 1995. But no one had counted on the political skill of Francis Everitt, the Stanford physicist who has been the project's principal investigator since 1981. Everitt was relentless in his lobbying, says one congressional staffer, targeting everyone from politicians representing his home state of California to members of federal appropriations committees. It worked. Every time NASA tried to pull the plug, the Congress resurrected the project.

Meanwhile, the probe's price tag kept growing. This was partly due to technical setbacks, but Stanford project managers also



What a blast: the probe has beaten the odds and is set to launch next year on a Delta II rocket.

complained that NASA was not providing enough cash to allow them to overcome problems without disrupting the schedule. For various reasons, the launch date kept slipping, and by this year the cost had ballooned to \$650 million.

Meanwhile, the study's scientific rationale became less compelling. Astronomers using the Rossi X-Ray Timing Explorer found indirect evidence of frame-dragging around a black hole in 1997, and a year later researchers reported a rough measurement of frame-dragging by Earth. This result, accurate only to within 20%, was based on data from the LAGEOS geodetic satellites, which are reflective spheres stationed in Earth's orbit — simple stuff by Gravity Probe B's exacting standards. But future LAGEOS studies may cut the margin of error to 5%, prompting the latest review of Gravity Probe B to admit "some erosion" in its scientific value in recent years.

After the Probe failed a ground test late last year (a minor problem caused by blown fuses, say its managers), an exasperated Ed Weiler, NASA's chief of space science, ordered yet another external review of the project's scientific merit and technical readiness. "I have no idea when and if it will ever be launched," Weiler complained at an academy meeting last March. He issued a threat to cancel the project

unless it passed the reviews. Once again, it did.

Over the years, the reviews have followed a pattern: acknowledge that the project is controversial and technically challenging, and then conclude — cautiously — that it should go ahead anyway. Val Fitch, a Princeton physicist and Nobel laureate who chaired the 1995 academy review, says that as so much money had already been spent on it, and its biggest technical hurdles had been overcome, it seemed wrong not to continue. But that opinion "was not unanimous by any means," he adds. Few on the panel thought that the results would deviate from Einstein's predictions, leading opponents to argue that there was no point in going ahead if the results could be predicted with such confidence.

Cost concerns

All that is now in the past. Even Weiler, who would gladly have killed Gravity Probe B years ago, says: "I look forward to a successful launch and a successful mission." His main concern now is preventing other projects from running so horribly over budget. It's about jumping on problems before it becomes impossible to turn back, he argues. "I'd much rather cancel a programme when we've spent a million dollars than when we've spent \$400 million," he says.

This new-found discipline is apparent in projects such as the Messenger mission to Mercury, scheduled for launch next May. When Messenger, which is part of NASA's Discovery Program of planetary missions and is managed by the Johns Hopkins University Applied Physics Laboratory, ran \$12 million over its \$300-million budget, Weiler considered cancelling it outright.

Instead, he asked his advisory council whether Messenger's scientific mission was valuable enough to allow it to expand slightly beyond its original cost. The council was not thrilled to be asked that question, Weiler says, because a 'yes' answer meant delaying funding for the next Discovery mission. Messenger was saved, but not without a stern warning to the scientific community on the importance of fiscal discipline.

Weiler says that NASA, too, has learned that its technical oversight must be beefed up to catch potential problems earlier. Discovery projects will be required to increase their budget reserve from 20% to 25% — meaning that, as Weiler admits, "there might be a little less science." And large projects, such as the successor to the Hubble Space Telescope (see *Nature* 422, 3; 2003), will spend more money on tech-

nological development and cost assessment, earlier in their journey to the launch pad.

Being aware of budget creep is one thing, but controlling it is another — particularly as NASA's projects become more technically difficult. "We have done the easy Discovery missions already," says Weiler. Pessimistic observers see a disaster looming, with an expensive effort to develop nuclear propulsion under way, the Mars Exploration Program overextended to the point that an ambitious 2009 Mars Rover mission already appears underfunded, and many smaller projects struggling to stay on time and on budget. Juggling all of these projects will be difficult, and some will almost inevitably be cancelled.

The European Space Agency, faced with similar problems but a smaller science budget, recently cast off ballast of its own. In response to "an outlook with no budget increase or other relief", the agency's Science Program Committee scrapped the proposed Eddington

mission to search for planets outside our Solar System and study the Sun's interior, and sacrificed the planned lander from its BepiColombo mission to explore Mercury.

Gravity Probe B will not be among NASA's cancellations. Weiler can take solace in the fact that, with the struggle nearly over, the mission will not eat up much more money. And who knows, maybe the science will equal the technical feats, and Gravity Probe B will discover something truly new.

"I've got my fingers crossed," says Eugene Parker, a veteran of the Gravity Probe B debate. He understands criticism of the project, but thinks it's a justified, even necessary, check of general relativity. Cosmology is taking off in many directions, from dark energy to string theory, all depending to some extent on Einstein's equa-

tions. Before heading deeper into unexplored territory, Parker says, "it would be nice to be absolutely sure, or surer than we are at present, of the general form of those equations".

Einstein didn't have to wait as long as NASA for his ideas to be tested. As early as 1919, astronomers measured the deflection of starlight during a solar eclipse — confirming a key prediction of general relativity. It is unlikely that Einstein had any doubt, however. In response to the news that his friend and fellow physicist Max Planck stayed up all night during the 1919 eclipse to see if it would confirm relativity, he said: "If he had really understood [the theory], he would have gone to bed the way I did." The chances are that Einstein would have stayed in bed for Gravity Probe B too. ■

Tony Reichhardt writes for *Nature* from Washington DC.



Political powers: Ed Weiler (top) and Francis Everitt.

KRT

R. RESSMEYER/CORBIS

STANFORD UNIV.

Filling the gap between knowing and doing

Research into delivery systems is needed to translate knowledge into improved health.

Sir — Your Editorial “In praise of Gates” (*Nature* 425, 435; 2003) urged the Bill & Melinda Gates Foundation to fund more upstream areas of basic research. I believe you are off the mark on this one.

The philanthropy of Bill Gates and its initial focus on neglected, high-burden diseases provide an important boost for the developing world. But I believe that the biggest gulf is between what we know and what we do in practice — the ‘know-do’ gap.

Although I would in no way deny the value of biomedical and upstream basic research, I believe we have placed too much emphasis on that and not enough on the translation of knowledge into actions to improve people’s health. Although programmes such as Roll Back Malaria and the Global Fund for TB, Malaria and AIDS are aimed at tackling these issues on the ground, they do not support research in this area. They need to consider the knowledge gaps that have to be addressed so that their programmes have the desired impact.

Research into public-health systems and

services and the contribution of the social and behavioural sciences (economics, anthropology, sociology and so on) are neglected fields. This is increasingly being recognized and addressed for specific diseases, but a more systematic approach is needed, which may benefit the entire health sector. Such research is critical in linking basic research to healthcare delivery, and in obtaining the participation and support of the people at whom new interventions are targeted.

For the Bill Gates billions to make a difference, a robust and efficient health system is needed in the countries where he hopes to have an impact. Without this, the multitude of new vaccines and drugs developed will have a negligible effect in reducing disease burdens.

To focus global attention on this neglected area, the World Health Organization will release a major report in 2004 entitled *Knowledge for Better Health*. The report’s release will be linked with a global Summit on Health Research, to be held in Mexico in November 2004. We hope to convene a high-level gathering of

leaders from the public, private and non-governmental sectors, the scientific community and the funders of health research to develop an action plan to address the key challenges of translating knowledge into health policy.

You referred to PPPs — public-private partnerships — as a successful means of bringing malaria drugs to market. No arguments there. But PPP could also stand for “publications, patents and professorships” — the mindset of many scientists in basic research.

It is my hope that the philanthropy of Bill Gates, through appropriate funding of research, will help to nurture another PPP mindset of “policy, practice and people”. He should put more of his money into research that will ensure that his laudable investments to date will have a real impact in improving the health and well-being of poor populations in the developing world.

Tikki Pang

Research Policy and Cooperation,
World Health Organization,
Avenue Appia, 1211 Geneva 27, Switzerland

Open access: the JCI has already shown it works

Sir — There has been much media attention surrounding the launch of the new open-access journal, *PLoS Biology* (see, for example, *Nature* 425, 554–555; 2003), which many scientists like myself hope will prove to be a success. However, some of the claims of originality for this venture are not warranted. The proposed model, in which authors pay the costs of free electronic access, has already been tested successfully by an existing high-profile periodical, the *Journal of Clinical Investigation* (JCI). Since the JCI went online (www.jci.org) in 1996, its content has remained completely free to everyone, with no barrier of any kind (J. B. Hawley *J. Clin. Invest.* 112, 968–969; 2003). This fits with the philosophy of its parent organization, the American Society for Clinical Investigation (ASCI).

Although the publishing and financial models of the JCI are not exactly the same as those of *PLoS Biology*, the only major difference appears to be that the ASCI holds copyright to articles. However, this is meant primarily to ensure the long-term integrity of the published work, and authors are not prevented from using the electronic version in any way that

would advance the cause of science.

Overall, the positive experience of the JCI with author-paid electronic free access augurs well for the future of *PLoS Biology*, and for other journals that may choose to take this path.

Of course, the same model may not apply easily to journals such as *Nature*, which also have to cover the costs of extensive front matter and general science policy and news reporting, thus serving a somewhat different role in scientific publishing.

Ajit Varki

Department of Medicine, University of California,
San Diego, 9500 Gilman Drive, La Jolla,
California 92093-0687, USA

Dr Varki has previously served as editor-in-chief of the JCI and president of the ASCI — Editor, Correspondence

Aircraft give a new view of jellyfish behaviour

Sir — In her recent News Feature “Close encounters of the jelly kind” (*Nature* 426, 12–15; 2003) Carina Dennis describes the use of submersibles to catalogue jellyfish in the ocean. However, it is not just in the inaccessible deep ocean that

technology is helping to improve our knowledge of jellyfish: the same applies to coastal areas. For example, scientists have struggled for more than half a century to record the distribution and abundance of giant *Rhizostoma* jellyfish in British and Irish waters. Today, low-flying aircraft are being used to survey wide areas.

This method revealed huge numbers of *Rhizostoma* — roughly one per square metre — in Carmarthen and Tremadoc bays in Wales from April to November this year. What drives these jellyfish to aggregate in these two large bays is unknown. However, both of these Welsh bays are known hotspots for sightings of jelly-feeding leatherback turtles, suggesting that some top predators are able to take advantage of this annual jelly-feast.

Graeme C. Hays*, Jon. D. R. Houghton*, Tom Doyle†, John Davenport†

*School of Biological Sciences,
University of Wales Swansea,

Singleton Park, Swansea SA2 8PP, UK

†Department of Zoology, Ecology and Plant Science,
University College Cork, Lee Maltings,
Prospect Row, Cork, Ireland

correspondence

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter. Published contributions are edited.

Fuzzy vision

A dark view of a future where the lines of self are blurred.

Tomorrow's People: How 21st Century Technology is Changing the Way We Think and Feel

by Susan Greenfield

Allen Lane: 2003, 304 pp. £20

Melvin Konner

It is best to be frank: when a boy from Brooklyn reviews a book by a baroness, there is, let us say, some trepidation. When the baroness is also a distinguished Oxford neuroscientist and director of the Royal Institution of Great Britain, self-confidence suffers further. Add to this the indisputable fact that (as the author points out) the reviewer is "hormonally challenged" by virtue of being male, and you have a serious set of obstacles. Nevertheless, one must try.

Susan Greenfield, author of *The Private Life of the Brain*, has here turned her skilled attention, intellectual power and literary gifts to the world to come. Baseball star Yogi Berra held that it is very difficult to make predictions, especially about the future. However, a large academic and industrial network worldwide is devoted to just such predictions. Trillions of euros are riding on them, and that prize, along with intrinsic human curiosity, underwrites the quest for believable visions of how we will live twenty, fifty or even a hundred years from now.

Greenfield's book is, first of all, a survey of the most plausible of those visions, sifted through the judgement of a wise and experienced scientist. If you want to know what sort of house you are likely to live in, how you will work, commute and communicate, and what you will do to monitor your health, you will find answers here that are the products of a rich imagination, but one informed by a broad knowledge of the cutting edge of technology and the aspirations of scientists and inventors working that frontier. The presentation is blessedly low on the gee-whiz, breathless style of much futurist writing, and avoids unreflective enthusiasm as carefully as it steers clear of luddite technophobia.

Some scenarios are both engaging and unsettling. Vast armies of nanorobots will do our bidding, while neuron-silicon interfaces blur the edge of the self. Computer chips will make every object we handle smart, adaptive, personalized and self-sustaining. Work will disappear into part-time efforts at home. Our bodies and their products will be automatically tested daily for hundreds of diseases, and the walls will light up with lucid explanations of signs, symptoms, diagnosis and treatment. We may relax in the evening by summoning a three-dimensional, virtual, life-sized sitcom that unfolds around us.



No hiding place: Tom Cruise tries to cover up his identity in the dystopian future of *Minority Report*.

Sex will be solitary, technologically assisted, varied enough to match our wildest fantasies, and unbelievably satisfying. Genes, of course, will be shaped at or before conception. Enhancement drugs will homogenize our minds, and generic, synthetic bacteria will be stored for a thousand uses determined by changing their genomes at the touch of a button — or, more likely, at a word. The first half of the book consists mainly of such revelations — some surprising, some familiar, but all thought-provoking.

The second half of the book is more important and more problematic. Here Greenfield deals with the human implications of these predicted advances. A chapter on education is sensible but fairly conventional; it demonstrates the impact of experience on the brain, suggesting that it will remain the best route to improving that organ. A clear discussion of consciousness and the brain brings the reader to the frontier of knowledge, but wisely does not attempt to solve the problem, instead simply intimating that consciousness itself will change in ways that have something to do with silicon.

In the remaining chapters on terrorism and human nature, followed by a brief concluding essay on our options, readers may become confused as to where Greenfield stands. She has declared herself no technophobe, but raises some ominous possibilities, and in the end she sounds an alarm. She quotes Bertrand Russell: "Science has not given men more self-control, more kindness, or more power of discounting their passions... Men's collective passions are mainly evil; far the strongest of them are hatred and rivalry directed toward other groups." And there is Freeman Dyson: "In the

long run, the central problem of an intelligent species is the problem of sanity." Greenfield does not seem to think that the science of the future will change human nature, and her opinion of our nature, though not as dark as Russell's, is not encouraging.

Science, she points out, will place unprecedented power in the hands of terrorists. Cyber-attacks could make processed food toxic by manipulating manufacturers' computers from half-way around the world. Bacteria could be designed to target the genetic vulnerabilities of a particular race. Hackers could bring whole nations to a halt. Human nature (along with its seven deadly sins) being in all likelihood unchanging, future science's brave new world could easily be far more dangerous than our own.

In the final pages, Greenfield considers the great majority of humanity who are likely to be left outside the new world. She writes of a science 'peace corps' that would somehow make this world and its advantages universal. But no less an authority on future technology than Bill Gates has doubts. At a conference on bringing technology to the developing world, he envisaged an African village with a computer: "The mothers are going to walk right up to that computer and say, 'My children are dying, what can you do?' Do you really have to put in computers to figure that out?" Gates has invested his vast philanthropic resources in vaccination and other basic public-health measures, not in 'siliconizing' Africa.

In the end, Greenfield writes: "The bottom line of this book is that the private ego is the most precious thing we each have, and it is far more vulnerable now than ever before." The ultimate priority, she continues, should be "not just the preservation, but also the

celebration of individuality". It ends with a warning: "Time could be running out" on our options, and "we may be the last generation of individuals able, or willing, to have them."

Greenfield almost seems, in the course of the book, to have turned into something of a technophobe. In any case, she leaves the reader feeling ambivalent about the wisdom of pushing technology rapidly, relentlessly and often rather thoughtlessly forwards. ■

Melvin Konner is in the Department of Anthropology, Emory University, Atlanta, Georgia 30322, USA. He is the author of The Tangled Wing: Biological Constraints on the Human Spirit, revised edition.

Jazzing up neuroscience

The Cognitive Neuroscience of Music

edited by Isabelle Peretz & Robert Zatorre
Oxford University Press: 2003. 466 pp.
£75, \$124.50 (hbk); £34.95, \$59.50 (pbk)

Petr Janata

From the soothing lyrical quality of lullabies to the driving pulse of techno, humans render sound patterns in myriad ways to serve diverse purposes. One category of human-generated sound patterns — speech — has the clear purpose of communication and has received considerable scientific attention. But the purpose of another immense range of sound patterns — those that underlie music — is manifold and mysterious. Compared with language, it has been difficult to ascribe to music a dominant role or specific survival value. Perhaps this is why there has been little effort to study the neural processes that underlie it.

The Cognitive Neuroscience of Music, a collection of chapters that originated from a meeting of the New York Academy of Sciences in May 2000, announces the arrival of music at the frontiers of neuroscience. Perhaps unsurprisingly, it begins by addressing allegations that searching for a biological basis for music is futile because music has no apparent adaptive role in human evolution. However, beyond this opening salvo and two chapters explicitly devoted to comparing music and language, the cognitive neuroscience of music is explored on its own merits.

The major strength of this multiauthored volume is that it captures the breadth of the field and introduces a diverse array of behavioural and cognitive neuroscience techniques, providing several perspectives on what music is and how it might be studied. For example, many authors use a bottom-up approach to focus on the neural correlates of the perception and production of individual tones (and their various qualities such as



Nice! But what happens to the brain when Pat Metheny hits the right notes?

pitch, timbre, loudness and duration), and ultimately on combinations or sequences of tones in various rhythmic patterns. Although they do not necessarily capture the essence of music, such studies of its building blocks are important because they help to identify how musical percepts arise and how they are ultimately differentiated from general multipurpose auditory perception.

Other approaches emphasize the importance of the higher-order structural relationships in music, such as scales, keys and musical phrases, to illustrate that even infants and non-musicians readily internalize such structures when they form memories and images of musical material. Yet other research programmes used expert musicians, with their highly specialized knowledge, in comparative studies of brain anatomy and functional organization, thereby examining an ethologically valid model of brain plasticity in humans.

Unfortunately, the benefits of these multiple viewpoints are diminished by the variation in presentation style. Some chapters, such as those that describe neurological disorders, musical imagery, the processing of expectations in music and language, and the evolutionary basis of music, are written for the lay reader and are a pleasure to read; other chapters have the density of conference proceedings aimed at the specialist. Although the chapters are grouped logically into several themes, the reader is left to stitch the information together across chapters and themes.

The heterogeneity that makes the reading difficult at times is akin to the free improvisation of a jazz ensemble before it revisits the recognizable and coherent theme at the head of a piece. Just as a jazz guitarist's improvisation follows a coherent path that may be different from the bass player's, the findings from the individual research programmes represented in this book are

intriguing and internally consistent, even though their interrelationships may not be obvious. For instance, there is no conceptual framework that reconciles observations of anatomical specificity, or 'modularity', of musical function (arising from neuropsychological assessments of lesion patients) with the variability in the networks that are recruited during musical tasks (as observed in functional neuroimaging experiments).

There has been a doubling in the number of music-related cognitive-neuroscience citations in the Science Citation Index since the conference was held that gave rise to this book. As the themes laid out here are revisited and elaborated on by the numerous players joining in the scientific jam session, the glorious moment will certainly come when all the themes blend into a harmonious whole. ■

Petr Janata is in the Department of Psychological and Brain Sciences, and at the Center for Cognitive Neuroscience, Dartmouth College, Hanover, New Hampshire 03755, USA.

Looking for a new home?

New Worlds in the Cosmos: The Discovery of Exoplanets

by Michel Mayor & Pierre-Yves Frei
(transl. Boud Roukema)
Cambridge University Press: 2003. 248 pp.
£18.95, \$30

Alan Boss

In 1995, Michel Mayor and Didier Queloz presented the first solid evidence for the existence outside our own Solar System of a planetary-mass body orbiting a Sun-like star. The number of likely extrasolar planets has since grown to more than 100, and both

NASA and its European equivalent, ESA, have plans to continue the search from space. Their ultimate goal is the discovery and characterization of Earth-like, habitable planets in orbit around nearby stars. In less than a decade the detection of extrasolar planets has gone from a field without a single credible example to one in which new detections are presented frequently, with a future that seems limitless. Within the next decade we will know how many Earth-like planets there are in the solar neighbourhood, and then we will have some idea about how many of them may support life.

Mayor and Frei present an account of this remarkable transformation that is personal, comprehensive and at all times understandable. They describe not only the recent successes of the field, but also its unfortunate history of failure, despite years of philosophical expectations that planetary systems should exist outside our own. The narrative sometimes slips into the first person as Mayor relates the details of the lifetime of astronomical adventure that led to his magnificent discovery. The book also describes in simple terms the physics behind planetary-system detection, and places it in the context of astronomy.

Detecting extrasolar planets is incredibly difficult, which is why it took until 1995 to find the first, despite the best efforts of many groups around the world. Even today, we have no direct evidence that extrasolar planets exist — we must rely on indirect evidence, primarily the wobble of a star around the centre of mass of a star–planet system, but also the dimming of a star's light as a planet in an edge-on orbit crosses in front of its star.

Nearly all of the detections made to date were achieved by the spectroscopic method, in which the star's wobble results in a periodic Doppler shift of its spectral lines. The wavelength shift is tiny, about one part in ten million at optical wavelengths, where the measurements are made. Arguably the first detection of a planetary system was made, also indirectly, by Alexander Wolszczan in 1992. He used radio-wave observations of a rapidly rotating neutron star to show that its wobble was caused by the presence of several Earth-mass objects. These 'pulsar planets' only whetted the appetites of those who were searching for planets around more benign stars, where life might have a chance to arise.

The goal of determining whether or not any of the Sun's neighbours might have an Earth-like planet in orbit is the dream that drives the field of extrasolar planet research. The discovery by Mayor and Queloz not only helped to launch this search but, along with the claims made for the ALH84001 martian meteorite in 1996, it sparked another new field of endeavour: astrobiology, the study of the origins, evolution, distribution and future of life in the Universe. As a result, astronomers now find themselves in serious

Exhibition

Seeing the light

Danish artist Jens Ferdinand Willumsen (1863–1958) is best known for his naturalistic paintings, only rarely turning his attention to scientific portraiture. But his 1913 portrait *En Fysiker*, shown here, is a striking, generic depiction of a young physicist. Seated on a stool, he leans forward, hands gripping his splayed knees, observing an experiment on electric-light rays.

It provides an intensely physical depiction of the researcher's absorption in his investigation of nature. His face and white coat are eerily lit by an incandescent tube, and a ceiling light and a glowing red lamp underneath the laboratory table provide additional light sources. The dramatic composition, the elongated, almost spiritual figure of the physicist, and the colours used all point to the influence of the Spanish artist El Greco on Willumsen's work.

The portrait was commissioned by engineer and industrialist Gustav Adolph Hagemann, who devoted his energy and cash to setting up an electrotechnical laboratory at the Polytechnical Institute in Copenhagen. He supported the 1903 Nobel prizewinner Neils Finzen in his investigations of the use of natural and artificial light in treating skin disease. In accepting the commission, Willumsen, by then a well-established artist, wrote to Hagemann that the picture "should have something to do with science and experimentation as suiting to your spirit". As executed, *En Fysiker* represents the new man of the twentieth century, exploring nature scientifically.



The portrait is included in a display of Willumsen's work in two rooms of the Statens Museum for Kunst in Copenhagen, which houses Denmark's national art collection. It is on loan until summer 2005, while the J. F. Willumsen Museum in Frederikssund is being rebuilt. Until then, it hangs alongside, and contrasts starkly with, *En Bjergbestigerske*, Willumsen's iconic painting of his wife, sculptor Edith Wessel, walking through a mountain landscape and enjoying the natural world without attempting to analyse it. **Colin Martin**

discussions with biologists about how an inhabited planet might appear, for example.

New Worlds in the Cosmos is the English-language translation of a French book that was published in 2001. Some of the book's content has been revised, but in such a rapidly changing field as extrasolar-planet detection it is impossible to stay completely up to date. Important news has been made, in terms both of scientific discoveries and of the ambitious space missions planned by ESA and NASA. The possible first discovery of an extrasolar planet by the transit method, OGLE-TR-56, may have been confirmed by follow-up spectroscopy. A planetary-mass object has been found to orbit a distant binary star in the ancient M4 globular cluster, implying that planet formation may have begun within a billion years of the Big Bang, some 13 billion years ago. If that's right, we may well be the late-comers to the galactic party.

All of the extrasolar planets discovered so far seem to be gas giants, similar to Jupiter or Saturn. Nearly all of them are in orbits

that are closer to the star than Jupiter is to our Sun, and some are orbiting at Earth's distance from the Sun — prohibiting the existence of an Earth-like planet in a similar orbit. Even 'hot Jupiters' that orbit much closer to the star would have passed through this orbit as they migrated closer, disrupting the formation of habitable worlds at more temperate distances.

In the next few years we can expect to learn how many nearby stars have gas giants in orbits that will allow a habitable Earth-like planet to exist. These stars will be the primary targets for a space telescope planned jointly by NASA and ESA that will be able to detect light from Earth-like planets directly. Through spectroscopy, this could tell us something about what it might be like to live there. Stay tuned for live news coverage, beginning in around 2015. Estate agents, get ready. ■

Alan Boss is in the Department of Terrestrial Magnetism, Carnegie Institution, 5241 Broad Branch Road NW, Washington DC 20015-1305, USA.

Four golden lessons

Steven Weinberg

When I received my undergraduate degree — about a hundred years ago — the physics literature seemed to me a vast, unexplored ocean, every part of which I had to chart before beginning any research of my own. How could I do anything without knowing everything that had already been done? Fortunately, in my first year of graduate school, I had the good luck to fall into the hands of senior physicists who insisted, over my anxious objections, that I must start doing research, and pick up what I needed to know as I went along. It was sink or swim. To my surprise, I found that this works. I managed to get a quick PhD — though when I got it I knew almost nothing about physics. But I did learn one big thing: that no one knows everything, and you don't have to.

Another lesson to be learned, to continue using my oceanographic metaphor, is that while you are swimming and not sinking you should aim for rough water. When I was teaching at the Massachusetts Institute of Technology in the late 1960s, a student told me that he wanted to go into general relativity rather than the area I was working on, elementary particle physics, because the principles of the former were well known, while the latter seemed like a mess to him. It struck me that he had just given a perfectly good reason for doing the opposite. Particle physics was an area where creative work could still be done. It really was a mess in the 1960s, but since that time the

work of many theoretical and experimental physicists has been able to sort it out, and put everything (well, almost everything) together in a beautiful theory known as the standard model. My advice is to go for the messes — that's where the action is.

My third piece of advice is probably the hardest to take. It is to forgive yourself for wasting time. Students are only asked to solve problems that their professors (unless unusually cruel) know to be solvable. In addition, it doesn't matter if the problems are scientifically important — they have to be solved to pass the course. But in the real world, it's very hard to know which problems are important, and you never know whether at a given moment in history a problem is solvable. At the beginning of the twentieth century, several leading physicists, including Lorentz and Abraham, were trying to work out a theory of the electron. This was partly in order to understand why all attempts to detect effects of Earth's motion through the ether had failed. We now know that they were working on the wrong problem. At that time, no one could have developed a successful theory of the electron, because quantum mechanics had not yet been discovered. It took the genius of Albert Einstein in 1905 to realize that the right problem on which to work was the effect of motion on measurements of space and time. This led him to the special theory of relativity. As you will never be sure which are the right problems to work on, most of the time that you spend in the laboratory or at your desk will be wasted. If you want to be creative, then you will have to get used

Scientist

Advice to students at the start of their scientific careers.

to spending most of your time not being creative, to being becalmed on the ocean of scientific knowledge.

Finally, learn something about the history of science, or at a minimum the history of your own branch of science. The least important reason for this is that the history may actually be of some use to you in your own scientific work. For instance, now and then scientists are hampered by believing one of the oversimplified models of science that have been proposed by philosophers from Francis Bacon to Thomas Kuhn and Karl Popper. The best antidote to the philosophy of science is a knowledge of the history of science.

More importantly, the history of science can make your work seem more worthwhile to you. As a scientist, you're probably not going to get rich. Your friends and relatives probably won't understand what you're doing. And if you work in a field like elementary particle physics, you won't even have the satisfaction of doing something that is immediately useful. But you can get great satisfaction by recognizing that your work in science is a part of history.

Look back 100 years, to 1903. How important is it now who was Prime Minister of Great Britain in 1903, or President of the United States? What stands out as really important is that at McGill University, Ernest Rutherford and Frederick Soddy were working out the nature of radioactivity. This work (of course!) had practical applications, but much more important were its cultural implications. The understanding of radioactivity allowed physicists to explain how the Sun and Earth's cores could still be hot after millions of years. In this way, it removed the last scientific objection to what many geologists and paleontologists thought was the great age of the Earth and the Sun. After this, Christians and Jews either had to give up belief in the literal truth of the Bible or resign themselves to intellectual irrelevance. This was just one step in a sequence of steps from Galileo through Newton and Darwin to the present that, time after time, has weakened the hold of religious dogmatism. Reading any newspaper nowadays is enough to show you that this work is not yet complete. But it is civilizing work, of which scientists are able to feel proud. ■

Steven Weinberg is in the Department of Physics, the University of Texas at Austin, Texas 78712, USA. This essay is based on a commencement talk given by the author at the Science Convocation at McGill University in June 2003.



Dive right in: exploring the unclear, uncharted areas of science can lead to creative work.

USANETWORKS/KOBAL COLLECTION/R. FOREMAN

Trees of life and of language

David B. Searls

The dating of ancient languages by a technique called glottochronology is undergoing a revival, stimulated by the computational and statistical methods used to tease out evolutionary relationships in biology.

This year marks a half-century since the publication¹ of a famous discovery that, by elucidating a remarkable mechanism for preserving and conveying information, led to a better understanding of our human heritage. This feat was performed by a partnership between an unconventional newcomer from a different field and a Cambridge scholar with specialist knowledge. Their insights depended on careful observations made by another academic, whose premature death from cancer ended her chances of sharing in the ultimate accolades.

The exploit described above is the decipherment of Linear B, an enigmatic scrawl found on clay tablets excavated at Mycenaean archaeological sites that proved to be an early example of script based on syllables, and a milestone on the path to alphabetic writing systems. An inventive young architect, Michael Ventris, joined forces with John Chadwick, a classicist with expertise in ancient dialects, to break the code and show (against the prevailing view) that Linear B was an early form of Greek. Another philologist, Alice Kober, had catalogued regularities dubbed 'triplets' in the texts that proved crucial to understanding

the inflectional structure of the underlying linguistic system. However, she disparaged Ventris's early efforts and didn't live to see the final breakthrough.

Beneath the surface similarities of this narrative to the discovery of the DNA double helix lie deeper connections that have long been observed between DNA and language, as well as between the evolution of species and of languages². Writing on page 435 of this issue, Gray and Atkinson³ bring an evolutionary biology mindset to the subject of historical linguistics. The issue they tackle is the origin of the Indo-European languages, a vast family of highly diverse tongues (see Fig. 1 of the paper, page 437). But at least as notable as their conclusions are their methods, which involve applying tools developed largely for reconstructing evolutionary relationships between species — phylogenies — using the information preserved in DNA.

Their work builds on foundations laid by the linguist Morris Swadesh, who, at just the time DNA and Linear B were being solved, was developing lexicostatistics, the quantitative study of vocabularies and vocabulary change. As a tool for inferring family trees of languages, lexicostatistics represented a departure from historical linguistic methods

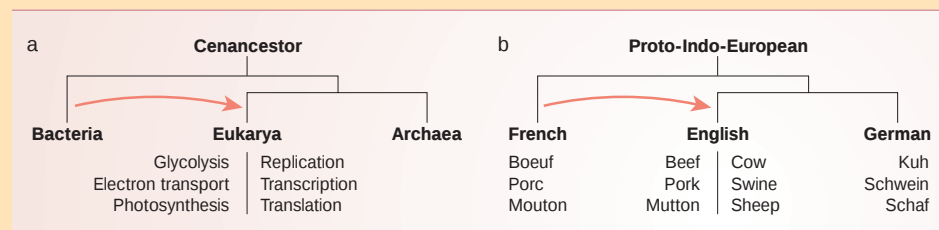
that characterized systematic, global patterns of change using comprehensive vocabularies and detailed knowledge of grammar and pronunciation. Swadesh focused instead on limited core vocabularies of 100–200 words (now called Swadesh lists), which reflect the most fundamental concepts expressed in any language and are expected to be relatively resistant to change. He then looked for similarities in the corresponding words in related languages, identifying the cognates (words presumed to derive from a common ancestor, or what in the realm of genes are called orthologues), thereby creating a straightforward and relatively objective metric of language kinship that made no attempt to account for the actual process of change. Simply put, two languages sharing 85% of cognates were judged to be more closely related than those sharing only 75%, and consequently to have split more recently.

Lexicostatistics was immediately controversial, particularly where Swadesh attempted to extend it to determine absolute times since language divergences by a method called glottochronology. This entailed determining an average rate of word substitution over time for recent language histories, and then extrapolating backwards. In postulating that

Box 1 Philology recapitulates phylogeny?

The figure here demonstrates similarities between reconstructions of language history and of the evolutionary relationships between organisms. Part **a** depicts the three domains of life — Bacteria, Eukarya and Archaea — in a commonly proposed topology rooted at a hypothetical 'cenancestor' (sometimes called the progenote or last universal common ancestor)^{5,7}. Uncertainty arises owing to extensive horizontal transfer of genes, as indicated by the arrow⁷. Eukarya (which include humans) may thus have derived some gene classes by descent and others by transfer.

The Eukarya and Archaea share similar versions of many 'informational' proteins, such as



those for the replication, transcription and translation of nucleic acids. But Eukarya appear to have derived a number of energy-related enzymes by gene transfer from Bacteria, probably by the wholesale incorporation of proteobacteria as so-called endosymbionts⁷ (a process in which the proteobacteria took up residence in the cells of early Eukarya, to mutual benefit).

In part **b**, French, English and German — three of the many languages thought to have arisen from an ancient proto-Indo-European tongue — are shown in their widely accepted evolutionary relationships. (See Fig. 1 on page 437 for a much more complete tree.) The core vocabulary of English is Germanic in origin. But extensive borrowings from the lineage leading to French, indicated by the arrow and

examples at the bottom, demonstrate how phylogenetic reconstruction of languages requires careful analysis of cognates, how word classes may behave differently (in this case, names of animals versus names of their flesh), and how rates might be expected to vary over time (for instance, by a kind of endosymbiosis after the Norman invasion of Britain in AD 1066).

D.B.S.

word 'half-lives' are intrinsic constants of language, Swadesh anticipated by more than a decade the notion of the 'molecular clock' that is central to sequence-based evolutionary biology².

In fact there are striking correspondences between the objections to Swadesh's methods and issues faced in phylogenetic reconstruction². Why should we believe that languages (or genomes) evolve at a constant rate over time, or that individual words and word classes (or proteins and protein families) have the same inherent rates as one another⁴? Do Swadesh lists (or core gene sets common to many genomes⁵) fairly represent overall evolutionary processes? How can we be sure that similarities reflect common ancestry, rather than chance convergences⁶? Conversely, how do we reliably recognize distant relatives whose spellings have drifted far apart (into what biologists call the 'twilight zone' of sequence similarity)? Why should we even presume that the 'tree of language' (or that of life) is a tree, as opposed to a sort of network, given that lexical borrowings and language mixture (and horizontal transfer between genomes⁷) are well-known occurrences? Box 1 provides some examples.

Over the years, historical linguists and evolutionary biologists have separately tackled such challenges with steadily increasing sophistication⁸, for instance supplanting Swadesh's overall similarity measure (which phylogeneticists would call a 'distance' method) with cladistic techniques that account for each word (or gene, or sequence residue, or any other observable character) to model the actual process of evolution⁶. Gray and Atkinson³ apply the latest computational tools — maximum-likelihood models and bayesian inference techniques, which offer a good framework for dealing with issues such as variable rates — to a data set of Indo-European languages developed and refined by lexicostatisticians⁴.

The topology of the resulting tree of languages holds no surprises for linguists, but the work goes further to estimate absolute ages with statistical support. Calibrating and cross-validating various branchings against known historical events, Gray and Atkinson develop robust confidence intervals for the date of the root of the tree, whence the Indo-European languages arose. This range appears to be several millennia too early to support a prominent theory that the proto-language was disseminated by nomadic Kurgan horsemen from the steppes of Asia, beginning about 6,000 years ago. But the dates fit well with the notion that Indo-European originated among nascent farming communities in Anatolia, in modern-day Turkey, some 2,000–4,000 years before that.

By breathing new life into glottochronology, this work will doubtless revive old debates. But at the same time it should stimulate even more cross-fertilization of ideas

among those studying the intertwined trees of life and language.

David B. Searls is in the Bioinformatics Division, Genetics Research, GlaxoSmithKline Pharmaceuticals, 709 Swedeland Road, PO Box 1539, King of Prussia, Pennsylvania 19406, USA.
e-mail: David_B_Searls@gsk.com

1. Ventris, M. & Chadwick, J. *J. Hellenic Stud.* 73, 84–103 (1953).
2. Searls, D. B. *Nature* 420, 211–217 (2002).

3. Gray, R. D. & Atkinson, Q. D. *Nature* 426, 435–439 (2003).
4. Kruskal, J. B., Dyen, I. & Black, P. in *Lexicostatistics in Genetic Linguistics* (ed. Dyen, A.) 30–55 (Mouton, The Hague, 1973).
5. Makarova, K. S. & Koonin, E. V. *Genome Biol.* 4, 115 (2003).
6. Ringe, D., Warnow, T. & Taylor, A. *Trans. Philol. Soc.* 100, 59–129 (2002).
7. Brown, J. R. *Nature Rev. Genet.* 4, 121–132 (2003).
8. Renfrew, C., McMahon, A. & Trask, L. (eds) *Time Depth in Historical Linguistics* (McDonald Inst. Archaeol. Res., Cambridge, UK, 2000).

Cell biology

Thanks for the memory

Jill C. Sible

In response to a transient hormonal cue, a developing egg commits irreversibly to a mature state. Surprisingly, this irreversible switch is composed of intrinsically reversible components.

Few decisions in life are truly irrevocable, even at the cellular level. Yet there are times when it's crucial not to turn back — during the development of multicellular organisms, for instance, when cells must make an irreversible commitment to a new fate in response to a transient stimulus. This process, called differentiation, is well known, but it's not really clear how a cell 'remembers' its commitment long after the signal, frequently a hormone, has disappeared. On page 460 of this issue, Xiong and Ferrell¹ show how brief exposure to the hormone progesterone triggers an irreversible switch in cell fate in developing frog eggs. The memory is sustained by positive feedback loops in the underlying control system. When the positive feedback is perturbed, the cell does the unthinkable and 'dedifferentiates', reverting to characteristics of the immature egg.

For more than 30 years, investigation of the maturation of frog eggs has been fertile ground for discoveries about how cells divide and differentiate. In 1971, Masui and Markert² demonstrated that a bit of cytoplasm taken from a mature frog egg and injected into an immature egg would induce maturation, replacing the need for progesterone. The authors named this mysterious activity MPF, for maturation-promoting factor. MPF was purified in 1988, and was shown to consist of two proteins — an enzyme called Cdc2, which attaches phosphate groups to other proteins, and an unstable co-factor called a cyclin, which is required for Cdc2 activity^{3,4}. A few years later, the activity of a second enzyme, MAPK, was shown to be required for egg maturation⁵. MAPK also phosphorylates proteins.

Subsequently, intensive biochemical and molecular investigations by many laboratories have produced a detailed picture of the signalling cascades that are elicited in an immature egg by progesterone. The picture has become quite complicated, with more

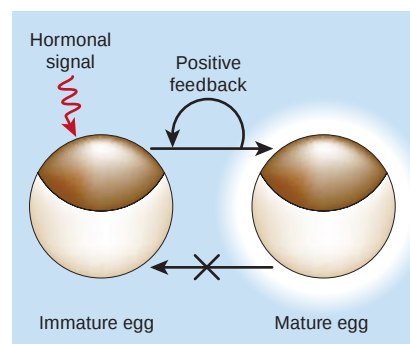


Figure 1 No turning back. When a hormone stimulates an egg cell to mature, the switch in cell fate is irreversible even after the hormonal signal is gone. Xiong and Ferrell¹ show that this long-term cellular memory is sustained by strong positive feedback in the underlying molecular control system.

than 30 different molecules involved. In fact, a recent comprehensive review of the literature⁶ required a poster-sized insert to depict all of the known biochemical responses of eggs to progesterone. In an attempt to sort through this web of molecular information, most scientists organize their thoughts around the two key enzymes — Cdc2 and MAPK — and the signalling events that activate each of them⁷.

Organizing the network in this way has helped investigators to appreciate the extensive positive feedback that is inherent in the molecular control system underlying egg maturation (see Fig. 1 on page 461). Positive feedback occurs when a molecule that is activated by a signalling pathway activates some earlier step in that pathway, thus strengthening and perpetuating the signal. Cdc2 activates itself in many ways: by inducing the synthesis of cyclin, by phosphorylating and turning off a Cdc2 inhibitor, and by phosphorylating and turning on a Cdc2 activator. Likewise, MAPK triggers the synthesis of Mos, a protein upstream in the

MAPK activation cascade. Furthermore, the Cdc2 and MAPK pathways activate each other. Except for the synthesis of a few key proteins such as Mos and cyclin, most of the signalling and feedback events in this system consist of phosphorylation⁶.

Despite this wealth of molecular information about egg maturation, and the appreciation of positive feedback in the molecular circuitry, until now no one could explain how eggs persist in the mature state — with high Cdc2 and MAPK activity — long after progesterone has been removed. Ferrell and Xiong previously carried out theoretical studies of the problem⁸, and found that positive feedback such as that seen in the Cdc2 and MAPK cascades could create a biochemical 'memory'. This prediction was certainly not intuitive, as these feedback events operate largely by phosphorylation — known to be rapid and readily reversible. Was it not more likely that a long-term response to a transient signal would be maintained by an irreversible event, such as the destruction of a key regulatory protein, maybe an inhibitor of Cdc2 or MAPK?

It appears not. Proof of Ferrell and Xiong's theoretical prediction required careful experimentation. In particular, the prediction demands that inhibiting the positive feedback will cause a mature egg to 'de-mature', at least with respect to key biochemical events. The authors now have compelling evidence to support this idea¹. They find that three different treatments that block the positive feedback between MAPK and Mos result in transient, rather than sustained, activation of both MAPK and Cdc2. So, positive feedback is necessary to maintain the biochemical memory of a mature egg (Fig. 1). Xiong and Ferrell also provide an additional, theoretical explanation for their data, with computational simulations that show how an essentially reversible switch can be rendered irreversible depending on the strength of the positive feedback in the system.

Beyond its particular contributions regarding the maturation of eggs, the work of Xiong and Ferrell¹ should have much broader impact owing to the approach they present for addressing questions of cellular function. As in work regarding the cellular switch that controls entry into, and exit from, nuclear division^{9,10}, Xiong and Ferrell's experiments were driven by theoretical predictions about the fundamental mechanisms that control such switches. These predictions derive from a systems-level view, but require quantitative, reductionist-style experimentation to test their validity. The ability to think and work on such a broad scale, from system through to molecule, is the most impressive aspect of Xiong and Ferrell's work. The notion of pairing theoretical and computational biology with experimental cell biology should catch on, and the positive feedback inherent in this interdisciplinary brand of

science is likely to drive many breakthroughs in our understanding of cellular controls. ■

Jill C. Sible is in the Biology Department, Virginia Polytechnic Institute and State University, Derring Hall, Blacksburg, Virginia 24061-0406, USA.
e-mail: siblej@v.edu

1. Xiong, W. & Ferrell, J. E. Jr *Nature* **426**, 460–465 (2003).
2. Masui, Y. & Markert, C. L. *J. Exp. Zool.* **177**, 129–145 (1971).

3. Gautier, J., Norbury, C. H., Lohka, M., Nurse, P. & Maller, J. L. *Cell* **54**, 433–439 (1988).
4. Gautier, J. et al. *Cell* **60**, 487–494 (1990).
5. Kosako, H., Gotoh, Y. & Nishida, E. *EMBO J.* **13**, 2131–2138 (1994).
6. Schmitt, A. & Nebreda, A. R. *J. Cell Sci.* **115**, 2457–2459 (2002).
7. Abrieu, A., Doree, M. & Fisher, D. J. *Cell Sci.* **114**, 257–267 (2001).
8. Ferrell, J. E. Jr & Xiong, W. *Chaos* **11**, 227–236 (2001).
9. Pomerening, J. R., Sontag, E. D. & Ferrell, J. E. Jr *Nature Cell Biol.* **5**, 346–351 (2003).
10. Sha, W. et al. *Proc. Natl Acad. Sci. USA* **100**, 975–980 (2003).

Planetary science

Conveyed to the Kuiper belt

Rodney Gomes

The small icy bodies that make up the Kuiper belt are the most distant objects known in the Solar System. A consistent picture is now emerging which suggests that these objects formed much closer to the Sun.

Since the first member of the Kuiper belt was discovered¹ in 1992, many unexpected features of their orbits and physical properties have been uncovered. One surprise was the very low total mass observed in the belt — about one-tenth of the mass of the Earth, when it was predicted to be a hundred times larger than that. To explain the missing mass, it has been proposed that collisions between Kuiper-belt objects over the lifetime of the Solar System have gradually transformed most of their mass into dust; bombarded by solar radiation, these dust grains are eventually expelled from the Solar System. But such theories have never been able to satisfactorily explain the extent of the depletion of the original

Kuiper-belt mass. On page 419 of this issue, Levison and Morbidelli² propose an alternative. They show that some objects that now exist in the Kuiper belt might have been pushed there from original positions near Neptune's present orbit: the original Kuiper-belt region could, in fact, have been virtually empty, and only a small amount of mass was subsequently deposited there.

According to current theory, the planets of the Solar System formed from a primordial disk of gas and dust, as the dust accumulated into gradually larger objects. Of course, in going from dust to planets there must have been intermediate stages, such as a disk of fledgling planets, or planetesimals, of roughly asteroid size. In regions where the total mass

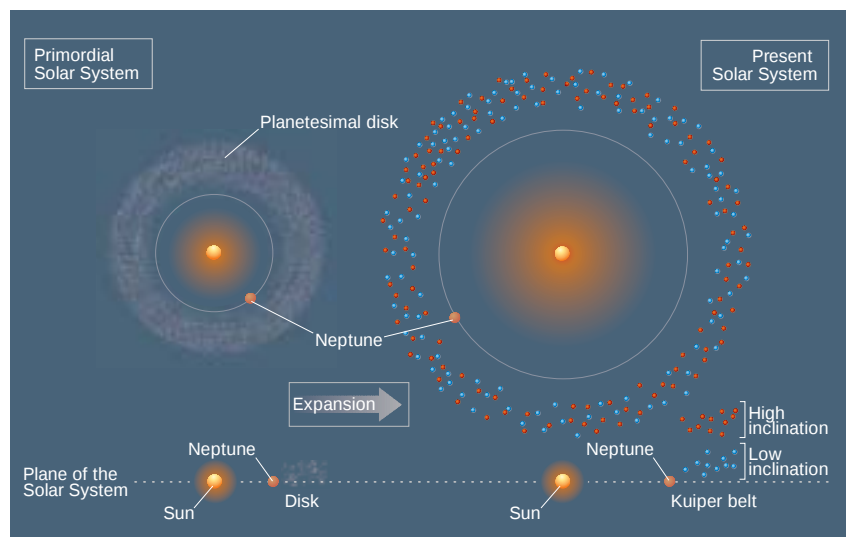


Figure 1 Expansion plan. The primordial, compact Solar System (left) was surrounded by a disk of asteroid-size planetesimals, which extended roughly as far as Neptune's present orbit. The gravitational pull of this disk eventually induced a planetary migration, during which the orbits of all the major planets (except Jupiter) expanded. As a consequence, most planetesimals experienced close encounters with the planets and were ejected from the Solar System. A few remnants were pushed out into stable orbits, forming the present-day Kuiper belt (right). Two distinct, yet complementary, mechanisms of gravitational interaction explain how these remnants became divided between high-inclination orbits³ and low-inclination orbits², with respect to the plane of the Solar System.

in the disk was not large enough for the accumulation process to produce planets, this planetesimal disk would have been preserved to the present day. This would certainly be the case for the Kuiper belt: because of its less dense mass distribution and large distance from the Sun, the accretion process in the belt would have ended when object sizes were no bigger than Pluto. However, in contrast to the present estimate for the Kuiper-belt mass, this accretion theory would require there to be around 10 Earth masses in the Kuiper-belt region, to allow the formation of objects as big as those presently seen in the belt.

The paucity of mass in the Kuiper belt is not the only enigma. Lying at such great distances from the rest of the Solar System, and experiencing no great perturbations from other large bodies, Kuiper-belt objects were expected to have orbits that were nearly circular and located close to the average plane of the Solar System. In fact, the orbits are quite eccentric, and are inclined out of the Solar System plane (Fig. 1). Planetary scientists have wondered how these orbits could have been so dynamically excited. My own idea³ is that these objects formed much closer to the Sun and were then propelled outwards by a mechanism involving close gravitational encounters with the outward-migrating, primordial Neptune. Other work⁴ shows that, if there had originally been a large mass beyond Neptune's present position, this planet would have moved much further out than it is today (effectively, into the Kuiper-belt region).

Thus, the puzzle seemed to be nearly solved. On the one hand, the original planetesimal disk would be truncated near Neptune's present position (Fig. 1), and was massive enough to form the large bodies now observed in the Kuiper belt; on the other hand, some objects would have been transported to the belt from this dense inner planetesimal disk by a mechanism that induced high-inclination orbits through close encounters with Neptune⁵. However, there was one piece that didn't fit. In addition to the Kuiper-belt objects in high-inclination orbits, there is a roughly equal number of objects at low inclination. These could not have been pushed out by the same mechanism.

Levison and Morbidelli² propose a solution. Their premise is based on another enigma — the fact that the Kuiper belt is considered to have an outer edge at about 7×10^9 km from the Sun (equivalent to 48 AU, or astronomical units). This distance is significant: at this point, the orbital periods of Kuiper-belt objects are twice that of Neptune, a feature known as the '1:2 mean motion resonance'. As Neptune migrated outwards through the primordial Solar System, it pushed out some of the objects in this resonance trap^{5,6}. This mechanism would naturally create orbit eccentricities, causing the resonant objects to approach close to Neptune and the other

major planets, such that they would eventually be ejected from the Solar System after a final gravitational encounter with Jupiter.

Levison and Morbidelli argue, however, that the influence of other, so-called secular resonances, inside the 1:2 resonance, would have kept the eccentricities and inclinations of some resonant objects low. A secular resonance is also based on commensurability of periods — not of the periods of the objects' motion around the Sun, but instead of the motion of the orbits themselves around the Sun. Such resonances are powerful, and can induce large variations in orbital eccentricities and inclinations, either raising or lowering them. According to Levison and Morbidelli's simulations, a small fraction of the resonance-trapped objects, once released through the rather jumpy migration of Neptune, would eventually be left in fairly low-inclination orbits in the Kuiper belt, owing to the influence of secular resonances. A few other objects remaining in the 1:2 resonance would set up the

outer edge of the Solar System as it is today.

Of course, this new set of ideas raises further questions. The main one is, how could the primordial Solar System be formed in a truncated disk? The few observations made of other planetary systems as they are forming indicate that they have radially expanded disks. Is the Solar System then a rare case? Regardless of the answer, what conditions could cause this truncation of a developing planetary system? This, I believe, will be a major topic in planetary science for years to come. ■

Rodney Gomes is at the Observatório do Valongo, Universidade Federal do Rio de Janeiro, and at the Observatório Nacional, Ladeira do Pedro Antonio 43, Rio de Janeiro 20080-090, Brazil.
e-mail: rodney@ov.ufrj.br

1. Luu, J. & Jewitt, D. *Nature* 362, 730–732 (1993).
2. Levison, H. F. & Morbidelli, A. *Nature* 426, 419–421 (2003).
3. Gomes, R. S. *Icarus* 161, 404–418 (2003).
4. Gomes, R. S., Morbidelli, A. & Levison, H. F. *Icarus* (in the press).
5. Malhotra, R. *Nature* 365, 819–821 (1993).
6. Malhotra, R. *Astron. J.* 110, 420–429 (1995).

Comparative genomics

Two worms are better than one

Mark Blaxter

The genome of the microscopic worm *Caenorhabditis briggsae* has been sequenced, and shows some remarkable differences from the genome of the better known — and physically similar — *C. elegans*.

In the early 1960s, when biologist Sydney Brenner was searching for a new model organism with which to study animal development and neurobiology, he screened a wide range of invertebrate species and chose the nematode *Caenorhabditis elegans* because it is easy to culture and transparent at all stages of its life cycle¹. This small worm is now famous, not least for being the first animal to have its whole genome sequenced². A close relative of *C. elegans* also passed by Brenner's microscope, and narrowly missed this accolade. This creature, *C. briggsae*, is physically very similar to *C. elegans* (it takes an expert to distinguish them), and the two have near-identical biology, even down to the minutiae of developmental processes. Surprisingly, however, their genomes are not so similar, as the sequencing of the *C. briggsae* genome to around 98% completion, reported in *Public Library of Science Biology*, now reveals³. Comparing the two species offers a new view of the patterns and processes that have shaped genomes, and raises many questions for the future.

From the first draft of the *C. elegans* genome², it was predicted that this microscopic worm has more than 19,000 protein-coding genes and 1,000 RNA-encoding genes. With the completion of the sequence to the last base pair (all 100,258,171 of

them⁴) in late 2002, these numbers have grown respectively to around 21,000 and 3,000. There is still vigorous debate as to how many of these genes are actually functional⁵, but what is clear is that the complexity of the *C. elegans* gene set contrasts markedly with the organism's morphological simplicity. For comparison, the more physically complicated fruitfly *Drosophila melanogaster* has only around 15,000 protein-coding genes⁶, and humans have some 40,000 (refs 7, 8).

The *C. briggsae* sequence reported by Stein *et al.*³, with its 19,500 protein-coding genes, provides comparative confirmation of most of the *C. elegans* gene set and, surprisingly, suggests that there may be another 1,300 *C. elegans* genes to add to the list. Stein *et al.* also propose more than 4,800 changes to current *C. elegans* gene predictions, such as the existence of new exons (the coding parts of genes, as opposed to their intervening, non-coding regions). These refinements will be crucial in exploiting this nematode as a model system. There are also some fascinating differences between the two species (why, for instance, does *C. elegans* have more than 700 chemoreceptor genes when *C. briggsae* gets by on just 430?), and many genes unique to each (about 800 per species).

Two other pairs of related genomes have been sequenced: humans^{7,8} and mice⁹ last

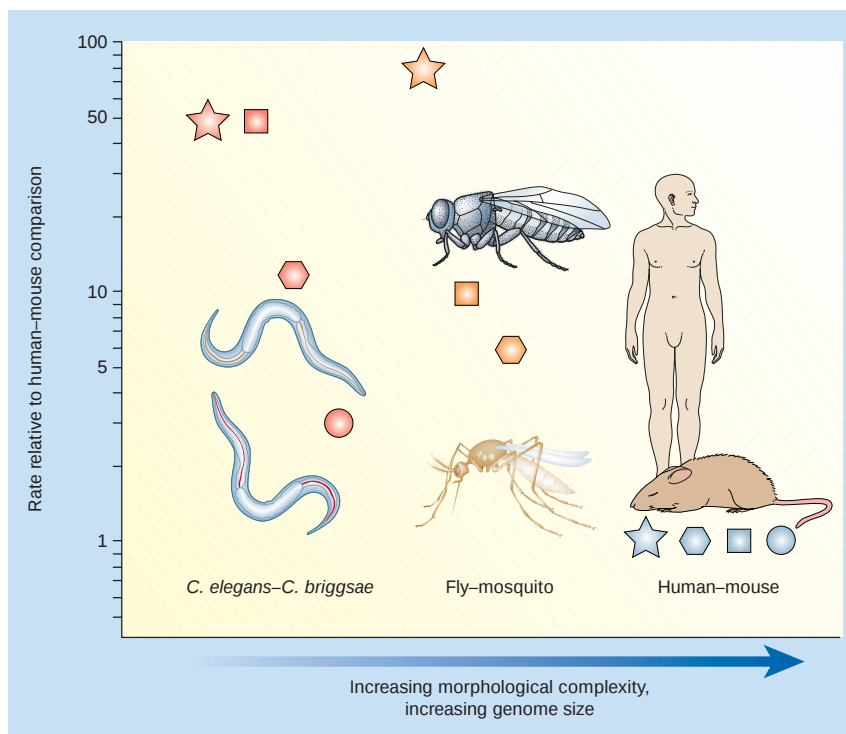


Figure 1 Rapid genome change and physical conservation in nematodes. Stein *et al.*³ have sequenced the genome of *Caenorhabditis briggsae*, and their comparison of its genome with that of *C. elegans* reveals rates of genomic change that stand in stark contrast to the lack of major morphological change that has occurred since the two species shared a common ancestor, around 100 million years ago. Humans and mice have undergone much more morphological evolution since they parted 85 million years ago, but have relatively more stable genomes. Flies and mosquitoes, separated by 250 million years, have an intermediate rate of change. The units on the y-axis are rates relative to the human–mouse divergence rates. Stars represent the rate of loss and gain of introns (non-coding gene regions); squares, the rate of genome reorganization; circles, the rate of ‘silent’ base-pair changes (not calculated for the fly–mosquito pair); hexagons, the number of blocks of genes whose order is conserved. The scale on the x-axis is arbitrary.

shared a common ancestor about 85 million years ago, and mosquitoes¹⁰ and fruitflies⁶ diverged around 250 million years ago. When did *C. briggsae* and *C. elegans* split? Judging from their morphology, one might think it was relatively recently, but the sequences tell a different story. Using equivalent genes from mosquitoes, humans and the two nematodes, Stein *et al.* estimate that the worms diverged between 80 million and 110 million years ago.

Do patterns of genome change help to describe the range of physical disparity between these various species pairs? The answer is a resounding no: the physically most similar pair, the nematodes, shows the most differences in terms of rate of genome evolution (Fig. 1). For instance, there are about three times more synonymous substitutions (‘silent’ base-pair changes that do not affect encoded proteins) between the two nematodes than there are between mice and humans. And changes in genome organization have occurred around 50 times as often.

Similarly, since the nematodes diverged, there have been about 0.5 changes in gene

structure — that is, in the pattern and spacing of exons — per gene. Since the divergence of mice and humans, there have been fewer than 0.01 changes in gene structure per gene⁹. Given all these changes, one question for the future is why the nematodes still look so similar. Stein *et al.* give a hint of an answer: they identify 1.3 million base-pair-level sequence matches between the two genomes, only a third of which correspond to coding portions of genes. The remaining sequence matches may represent conserved control elements that coordinate gene expression to produce physically similar organisms.

Another interesting finding comes from a look at the pattern of gene evolution along chromosomes. In *C. elegans*, the ‘arms’ of the chromosomes were found to be rich in repeated sequences and genes that have no similarity to those of other organisms, and undergo frequent genetic exchange². By contrast, the centres of the chromosomes had few repeats and contained more genes that are also found in other animals. Comparison with *C. briggsae* reinforces this model: genes on the arms are significantly more different to genes in other organisms than are those in

the centres, and gene order is less likely to have been preserved on the arms. This is strikingly reminiscent of the linear chromosomes of streptomycete bacteria, where exotic functions, such as antibiotic synthesis, are encoded on the arms and housekeeping genes are encoded in the centre. In *C. elegans*, gene-knockout studies have identified blocks of genes from the same chromosome that are expressed in the same tissues or stages of the life cycle¹¹. How these blocks are maintained in the face of randomizing genome reorganization remains unknown.

Rapid change is not the rule, however. Despite having undergone more than 4,000 chromosomal breakages since they parted³, *C. briggsae* and *C. elegans* have the same number of chromosomes: most rearrangements occur within chromosomes rather than between different ones. This pattern may be generally true of nematodes, as comparisons with the distantly related human parasite *Brugia malayi* also suggest a preponderance of intrachromosomal rearrangements¹². Moreover, the nematode group to which *C. elegans* and *C. briggsae* belong, the Rhabditida, has a remarkable constancy of chromosome number, with six or seven chromosomes being the norm¹³. So another question for the future is how this set number of chromosomes is maintained.

The publication of the *C. briggsae* genome sequence will undoubtedly spur many workers to use this species in comparative work, and a programme of identifying ‘true’ genes (rather than ‘predicted’ ones) by mutating them is already under way¹⁴. A third nematode genome, that of *B. malayi*, should be added soon¹⁵, and plans are also afoot to fill out the caenorhabditid tree with genome sequences of related species, such as that of *C. remanei*, approximately equally related to the two with sequenced genomes. These additional genomes will nourish comparative genomics, and bolster *C. elegans*’ position as an invaluable model organism. ■

Mark Blaxter is in the School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK. e-mail: mark.blaxter@ed.ac.uk

- Wood, W. B. (ed.) *The Nematode Caenorhabditis elegans* (Cold Spring Harbor Lab. Press, New York, 1988).
- The *C. elegans* Genome Sequencing Consortium *Science* **282**, 2012–2018 (1998).
- Stein, L. D. *et al.* *PLoS Biol.* **1**, 166–192 (2003).
- www.wormbase.org
- Mounsey, A., Bauer, P. & Hope, I. A. *Genome Res.* **12**, 770–775 (2002).
- Celniker, S. E. *et al.* *Genome Biol.* **3**, research0079.1–0079.14 (2002).
- International Human Genome Sequencing Consortium *Nature* **409**, 860–921 (2001).
- Venter, J. C. *et al.* *Science* **291**, 1304–1351 (2001).
- Mouse Genome Sequencing Consortium *Nature* **420**, 520–562 (2002).
- Holt, R. A. *et al.* *Science* **298**, 129–149 (2002).
- Kamath, R. S. *et al.* *Nature* **421**, 231–237 (2003).
- Guiliano, D. B. *et al.* *Genome Biol.* **3**, research0057.1–0057.14 (2002).
- Blaxter, M. L. *Int. J. Parasitol.* **30**, 347–355 (2000).
- http://wormlab.caltech.edu/briggsae/index.html
- www.tigr.org/tdb/e2k1/bma1

Astronomy

Beacons in the distant cosmos

Luigi Piro

Gamma-ray bursts are the brightest and most distant explosions in the Universe. A convincing body of evidence now links these bursts to supernovae, but there is still more to learn about their origins.

Although previously limited to a small pool of specialists, research on γ -ray bursts (GRBs) has now entered the mainstream of astrophysics, as attendance at two recent conferences* attests. Most progress in the field concerns the most numerous class of GRBs, the so-called long bursts, which last more than two seconds. The brief, paroxysmic emission of γ -rays in a long burst is followed by a long-wavelength, fainter afterglow over much longer timescales^{1–3}. We know that the bursts take place in galaxies at cosmological distances⁴, and they are probably associated with the birth of a black hole following the collapse and explosion of very massive stars. The basic mechanism producing the GRB and its associated afterglow is understood in terms of an energetic, collimated flow of matter, called the fireball model⁵. But what in turn generated that flow of matter; what is the progenitor of a GRB?

The nature of the progenitor was a focal issue at both conferences. The quest is not easy. The radiation observed in a GRB and its afterglow are produced by shock waves at considerable distances from the central source; the fireball brings with it very little memory of the central engine, depending only on basic parameters such as the total energy or the collimation angle of the flow (T. Piran, Hebrew Univ., Jerusalem). However, by examining the environment surrounding GRBs, astronomers have found some clues.

The connection between long GRBs and certain types of supernova was first suggested by the coincidence of a burst on 25 April 1998 with a supernova (SN1998bw)^{6,7}. A later recovery in the brightness of some GRBs has also been attributed to an underlying supernova component⁸, but this was not unambiguous proof. However, a bright burst detected on 29 March this year showed, coinciding with re-brightening, broad spectral features that were basically identical to those observed in SN1998bw (J. Hjorth, Univ. Copenhagen). This burst is thus firmly linked to a coincident supernova, SN2003dh.

Radiation from particular atoms — iron⁹, silicon, sulphur and magnesium — seen in the X-rays associated with some GRBs is a tell-tale sign of a supernova explosion (J. Osborne, Univ. Leicester; D. Watson,

Univ. Copenhagen; N. Butler, Massachusetts Inst. Technology). All metals (astronomically speaking, anything heavier than helium) in the Universe are the end result of nucleosynthesis in stars, and are ejected in supernovae. Data also indicate outflowing velocities around GRBs that are consistent with supernova ejecta. Each single measurement is not of overwhelming statistical significance, but I find the body of data rather compelling. In one scheme of events (D. Lazzati, Inst. Astronomy, Cambridge), atomic emission lines are produced when material ejected by the supernova is illuminated by the X-ray component of the GRB. This requires a two-step sequence, with the supernova occurring at least a month before the GRB, as in the 'supranova' model¹⁰; in the aftermath of a supernova, a rapidly rotating neutron star is formed, which later collapses into a black hole, powering the GRB explosion.

But the GRB and supernova explosion around 29 March occurred within days of each other. This near simultaneity is more consistent with the 'collapsar' model (S. Woosley, Univ. California at Santa Cruz). It postulates that the core of a massive rotating star collapses into a black hole, and the huge amount of energy released accelerates material along the rotation axis, creating an energetic jet that expands out of the stellar surface. This jet eventually produces the GRB and its afterglow. A substantial fraction of energy is also dumped in the stellar interior, and this drives the supernova explosion. Radiation at optical wavelengths from this so-called engine-driven supernova is generated through the production of nickel (⁵⁶Ni) atoms and their subsequent decay into iron atoms (⁵⁶Fe). But, depending on the precise conditions, it may be that stable iron isotopes are synthesized in preference to ⁵⁶Ni, with the result that the supernova explosion is not visible at optical wavelengths (A. MacFayden, Caltech). This could explain the apparent dichotomy, admittedly based on sparse data, between iron X-ray features and optical supernova signatures in GRBs: usually only one or the other is found. The X-ray lines could be produced near the central engine, powered by the energy wasted as the jet pierces through the stellar interior (P. Mészáros, Pennsylvania State Univ.).

A newly discovered type of cosmic explosion, called X-ray flashes (XRFs), has generated considerable interest. These resemble GRBs in almost every respect¹¹, and they



100 YEARS AGO

A correspondent of the *Times* directs attention to the wise recognition given to science and other branches of learning by Continental nations on all occasions of national importance; and the comparison he makes with our own official customs is not creditable to our dignity. When a monarch or the supreme authority of a State visits the Court of another nation, men of "light and leading" are usually invited as guests to meet him. These are the men who give distinction to a nation; and a people which fosters intellectual accomplishments cannot conceive a State function in which they are not represented. Here, however, there is little pride in the glory which learning brings to a State, and little encouragement is given to the men who devote themselves to the advance of knowledge.

From *Nature* 26 November 1903.

50 YEARS AGO

Dr. L. G. C. E. Pugh opened his talk on nutritional aspects of Himalayan mountaineering by a description of the experimental work which has led to the belief that a high carbohydrate diet increases a man's tolerance to great altitudes, and supported this by the experiences of people on earlier expeditions to great heights. In earlier Everest expeditions, climbers took an enormous store of luxury foods such as chicken in aspic; but they found when they got high up that they wanted something more ordinary, and often relied mainly on sugar and other carbohydrate foods. In general, diets have tended to get simpler and simpler, and local foods have been used extensively on some expeditions. In Dr. Pugh's opinion, however, curry and tea were overdone on the Cho Oyu expedition! He emphasized the importance of labelling the packages of food clearly so that the climbers do not find themselves with nothing but potatoes to eat for several days on end — as once happened, and which he thought, with good reason, did not improve morale. On the recent Everest expedition, the average calorie consumption was 4,300 a day during the approach march and 3,900 at 19–22,000 ft... Special fancies were catered for, however, and on the final assault Hillary chose to eat sardines and biscuits.

From *Nature* 28 November 1953.

*Gamma-ray Bursts: An Enigma and a Tool, Santorini, Greece, 24–27 August 2003; GRB 2003, Santa Fe, New Mexico, USA, 8–12 September 2003.

constitute at least one-third of the total population of GRBs. With the measurement of the distance to two XRFs, it has been shown (D. Lamb, Univ. Chicago) that these events follow the same relationship between isotropic energy and peak energy as do GRBs¹² — which strongly suggests that GRBs and XRFs share the same origin. One possibility is that an XRF is a GRB seen off-axis, rather than head-on. In the collapsar model, in addition to the collimated, high-energy jet that gives rise to the GRB, a considerable fraction of energy is ejected at wider angles and could produce a burst that peaks at X-ray wavelengths, rather than in γ -rays.

The evidence in favour of a connection between supernovae and GRBs — and probably also XRFs — seems overwhelming. But there are still some inconsistencies in the picture: a radio survey of a particular sample of supernovae (called type Ib and Ic) thought to be connected with GRBs shows that basically none of these objects has properties similar to those of GRB afterglows or of SN1998bw, the first supernova to be linked to a GRB (E. Berger, Caltech); this makes it difficult to understand how there can be simple underlying physics across the board for GRBs and supernovae. And despite the efforts made to assemble the observational evidence into a single model, not all the pieces of the puzzle fit. For instance, the observed evolution of most of the afterglows is awkward to explain in the collapsar model (R. Chevalier, Univ. Virginia), although not a problem for the supranova model (A. Konigl, Univ. Chicago).

GRBs are among the most distant sources that we can observe in the Universe, because they are so bright. For cosmology, this opens the exciting perspective of GRBs as probes of the unobserved history of the Universe, looking back to when the first population of stars and primordial galaxies formed (G. Djorgovsky, Caltech). Present estimates indicate that although a substantial fraction of GRBs actually lie in this region of the Universe (technically, beyond a 'redshift' of five; A. Loeb, Harvard Univ.), they are not visible at optical wavelengths because this radiation is absorbed by hydrogen in the intergalactic medium. Interestingly, several GRBs have no detected optical counterpart¹³. And because most distance (redshift) measurements are derived from optical data, analyses are at present biased against the most distant GRBs: X-ray and infrared measurements are needed as well.

The SWIFT observatory, launching in May 2004, will deliver hundreds of precise, fast localizations of GRBs at γ -ray wavelengths and track within minutes their X-ray and optical afterglows (N. Gehrels, NASA Goddard Space Flight Center), improving considerably on the pioneering method first used by the BeppoSAX mission (F. Frontera, Univ. Ferrara). In addition, the HETE-2 satellite, currently in orbit, should continue

to provide complementary X-ray localizations of GRBs (G. Ricker, Massachusetts Inst. Technology). For the study of GRBs, the future is certainly bright.

Luigi Piro is at the Istituto Astrofisica Spaziale Fisica Cosmica, INAF, Via Fosso del Cavaliere, I-00133 Rome, Italy.

e-mail: piro@rm.iasf.cnr.it

- Costa, E. *et al.* *Nature* **387**, 783–785 (1997).
- van Paradijs, J. *et al.* *Nature* **386**, 686–689 (1997).

- Frail, D. A. *et al.* *Nature* **389**, 261–263 (1997).
- Metzger, M. R. *et al.* *Nature* **387**, 878–880 (1997).
- Mészáros, P. *Annu. Rev. Astron. Astrophys.* **40**, 137–169 (2002).
- Pian, E. *et al.* *Astrophys. J.* **536**, 778–787 (2000).
- Galama, T. J. *et al.* *Nature* **395**, 670–672 (1998).
- Bloom, J. S. *et al.* *Nature* **401**, 453–456 (1999).
- Piro, L. *et al.* *Science* **290**, 955–958 (2000).
- Vietri, M. & Stella, L. *Astrophys. J.* **507**, L45–L48 (1998).
- Heise, J. *et al.* in *Gamma-Ray Bursts in the Afterglow Era* (eds Costa, E., Frontera, F. & Hjorth, J.) 16–21 (Springer, Berlin, 2001).
- Amati, L. *et al.* *Astron. Astrophys.* **390**, 81–89 (2003).
- de Pasquale, M. *et al.* *Astrophys. J.* **592**, 1018–1024 (2003).

Structural biology

Dual approach to a light problem

Werner Kühlbrandt

The structure of the last of the major pigment-containing protein complexes involved in photosynthesis is now revealed. The details complete our picture of electron shuttling in this vital process.

It is not uncommon for notable scientific progress to be made simultaneously by two independent teams of researchers. The latest example concerns the long-awaited structure of the cytochrome b_6f complex, described by Stroebel *et al.* on page 413 of this issue¹ and by Kurisu *et al.* in *Science*². Cytochrome b_6f mediates the flow of electrons between photosystems II and I in the photosynthetic membranes of plants and cyanobacteria. It is the last of these large, pigment-containing protein complexes to yield to detailed crystallographic analysis,

and the structures offer some surprising insights into how it works.

Stroebel and colleagues¹ examined the b_6f complex found in the chloroplasts — the photosynthetic organelles — of a unicellular alga¹, whereas Kurisu and co-workers² studied the same complex from a cyanobacterium. Both complexes have essentially the same structure. This in itself is astonishing, given that the two types of organisms are separated by an evolutionary distance of roughly 1,000 million years.

Plants and cyanobacteria have the unique

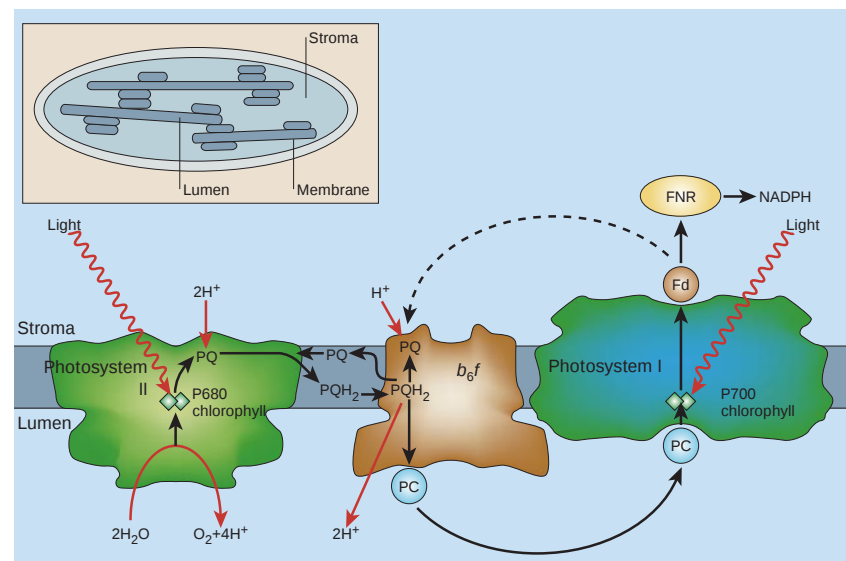


Figure 1 Electron transport in oxygenic photosynthesis. Photosystem II uses solar energy to withdraw electrons from water, generating oxygen as a waste product. Two electrons are accepted by plastoquinone (PQ), which then binds two protons for electroneutrality. The reduced plastoquinone (PQH₂) diffuses in the membrane to the cytochrome b_6f complex. One electron is transferred to plastocyanin (PC); the other passes to another PQ molecule. The PC diffuses to photosystem I, which uses solar energy to propel the electron against a potential gradient across the membrane. The electron is accepted by ferredoxin (Fd) and transferred to an enzyme (ferredoxin:NADP⁺ reductase, FNR) that converts NADP into NADPH, the primary product of photosynthesis. X-ray structures of the cytochrome b_6f complex^{1,2} suggest that Fd can also deliver its electrons back to b_6f in a cyclic electron flow (dashed line). Black arrows, electron transfer; red arrows, proton transfer. The inset shows a chloroplast.

ability to use solar energy to withdraw electrons from water, the most abundant substrate on Earth. Unlike the purple and green bacteria, which use more energy-rich substrates, plants and cyanobacteria need two big protein complexes — the photosystems — to bridge the large energy gap between water and the stable reducing agent NADPH. This agent enables the synthesis of organic substances such as sugars and starch.

Photosystem II is the complex that uses solar energy to withdraw electrons from water, generating oxygen as a waste product (Fig. 1). The immediately useful product of this photosystem is the unstable, reduced form of plastoquinone, a lipid-soluble molecule that has been endowed with two electrons by photosystem II. The cytochrome b_6f complex recycles this reduced plastoquinone, stripping it of its electrons and releasing two protons that contribute to the electrochemical gradient across the photosynthetic membrane. This gradient is ultimately used to make ATP, the main energy-storing molecule in living cells. The oxidized plastoquinone goes back to photosystem II to be reduced again. Meanwhile, the b_6f complex feeds one of the stripped electrons into photosystem I, which uses solar power to transfer it to the opposite membrane surface for NADPH synthesis. The other electron passes through the b_6f complex and reduces another plastoquinone molecule.

In addition to this linear mode of electron flow, the system can switch to a cyclic mode, in which photosystem I returns some of its electrons to the b_6f complex rather than feeding them into biosynthesis. This is necessary to balance the activities of the two photosystems, which depend on the variable amounts of solar energy absorbed by each. The cyclic mode of electron flow is poorly understood, but the X-ray structures of the b_6f complex^{1,2} suggest a plausible mechanism.

Cytochrome b_6f comprises several membrane-spanning proteins, namely cytochrome b_6 , cytochrome f , and a protein that harbours an iron-sulphur cluster. The complex owes its deep brown colour to pigment molecules attached to the proteins; these include four haems, a chlorophyll and a β -carotene. Stroebel *et al.* and Kurisu *et al.* now find that the membrane-embedded part of the b_6f complex has an extensive hydrophobic cavity, which is partly filled with membrane lipids, but allows access to the two plastoquinone-binding sites of the complex.

As expected from knowledge of a related structure, the mitochondrial cytochrome bc_1 complex^{3–6}, the two b -type haems are located next to the plastoquinone-binding sites (Fig. 2). Unexpectedly, however, one of these sites contains an extra haem group, not present in the bc_1 complex. This pigment (which Stroebel *et al.* call haem c_i and Kurisu *et al.* call haem x) is covalently attached to the

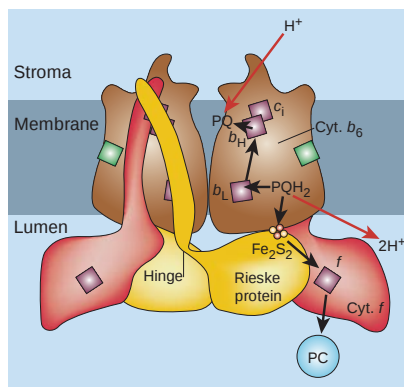


Figure 2 Electron-transfer routes in the cytochrome b_6f complex, incorporating the new findings^{1,2}. The complex is a dimer, with two identical sets of components. At the PQH₂-binding site, two protons are released on the luminal side of the membrane. The iron-sulphur cluster (Fe₂S₂) takes one electron from PQH₂ and passes it to haem f in cytochrome f , where it is picked up by PC. The Fe₂S₂ cluster is attached to an iron-sulphur protein (Rieske protein; ISP in ref. 2), which moves on a hinge to bridge the gap between haem b_L and f . The second electron passes via haem b_L and b_H (b_H and b_L in ref. 2) to a PQ bound at a site near the stromal membrane surface. PQ binds a proton, adding to the pH gradient across the membrane. The newly discovered haem c_i (x in ref. 2) is well placed for the re-uptake of electrons in cyclic electron flow. Purple squares, haems; green squares, chlorophylls; black arrows, electron transfer; red arrows, proton transfer.

cytochrome b_6 polypeptide — making it all the more astonishing that it has escaped discovery until now. Given its position, the most likely function of this extra haem is in cyclic electron flow (Fig. 1). The X-ray structures now make it possible to design experiments to study this elusive process in detail.

Together, the haems mark the path that electrons take between the two plastoquinone-binding sites. The other path to the surface-exposed cytochrome f leads via an iron-sulphur cluster, which is found in a protein domain that moves on a hinge to deliver the electron to haem f (Fig. 2). This movement has been inferred from the different positions of the analogous iron-sulphur domains in bc_1 complexes^{3–6}. Kurisu and colleagues' structure² provides evidence for the same mechanism in the b_6f complex.

The new haem is not the only surprising feature of the b_6f complex. Even though their function in the complex was not clear, it was widely expected that the chlorophyll and carotene would be in close contact, allowing the carotene to defuse the chlorophyll — excitation of an isolated chlorophyll molecule by light would produce oxygen radicals. But both structures show that the two pigments are too far apart to prevent this

potentially lethal reaction. It was also hoped that the structure would explain why the chlorophyll is there, but its role remains a mystery. Kurisu *et al.*² suggest that it may simply be a space filler (although it would be a highly dangerous one!), whereas Stroebel *et al.*¹ propose that it acts as a sensor in the interaction with photosystem I — which might then provide the absent carotene. Such large functional assemblies, or 'super-complexes', of cytochrome b_6f and the photosystems have been postulated^{7,8}, and might indeed exist in the crowded photosynthetic membrane, but they have not yet been seen.

Both teams^{1,2} have been working towards a high-resolution structure of the b_6f complex for well over ten years. Why did it take so long? Crystallizing membrane proteins is nearly always problematic, because detergents are required, which often make it difficult to grow diffraction-quality crystals. To make matters worse, the b_6f complex is easily disrupted once it is removed from the photosynthetic membrane. Each team came up with a different ruse to get around this problem. Stroebel *et al.*¹ engineered a tag into the complex, using it to get hold of the complex and purify it quickly — so avoiding prolonged exposure to damaging detergents. Kurisu *et al.*² chose the inherently more stable complex from a thermophilic organism, but found that they needed to add back lipids removed during purification before they could obtain well-ordered crystals⁹.

The structure of cytochrome b_6f completes our picture of the molecular events that produce the oxygen in the air we breathe. This is one of the first membrane processes that we can follow in its entirety. In the future, we can hope to understand many other, no less fundamental and fascinating membrane systems, including those that enable us to see, hear, taste and think. For all of these we will need information on the structure of membrane proteins at similarly high resolution. But this information cannot be obtained without similar high-risk, long-term efforts.

Werner Kühlbrandt is at the Max Planck Institute of Biophysics, Marie-Curie-Strasse 15, D-60439 Frankfurt am Main, Germany. e-mail: werner.kuehlbrandt@mpibp-frankfurt.mpg.de

1. Stroebel, D., Choquet, Y., Popot, J.-L. & Picot, D. *Nature* 426, 413–418 (2003).
2. Kurisu, G., Zhang, H., Smith, J. L. & Cramer, W. A. *Science* 302, 1009–1014 (2003).
3. Xia, D. *et al.* *Science* 277, 60–66 (1997).
4. Zhang, Z. *et al.* *Nature* 392, 677–684 (1998).
5. Iwata, S. *et al.* *Science* 281, 64–71 (1998).
6. Hunte, C. *et al.* *Structure Folding Design* 8, 669–684 (2000).
7. Zhang, H. *et al.* *J. Biol. Chem.* 276, 38159–38165 (2001).
8. Joliot, P. & Joliot, A. *Proc. Natl Acad. Sci. USA* 99, 10209–10214 (2002).
9. Zhang, H., Kurisu, G., Smith, J. L. & Cramer, W. A. *Proc. Natl Acad. Sci. USA* 100, 5160–5163 (2003).

Earth science

The ultraslow difference

Jason Phipps Morgan

The speed at which mid-ocean ridges grind out new ocean floor varies considerably. The slowest-spreading ridges are especially tough to study — but the latest data show that they are especially intriguing.

Earth exploration still has the power to astound. Mid-ocean ridges, also called spreading centres, produce new ocean crust at different rates and are classified accordingly. Dick and colleagues (page 405 of this issue¹) show that ridges exhibit fascinating behaviour when the spreading rate is 'ultraslow', less than about 1.2 cm yr^{-1} . The ridges they have investigated, the Gakkel and southwest Indian ridges, have long, quiescent stretches and oblique, non-volcanic, spreading segments that further differentiate them from their faster-spreading relatives.

Mid-ocean ridges were first intensively studied in the late 1950s, when it became evident that the Mid-Atlantic Ridge has the same form as the coastlines of South America, North America and Africa — almost as if all of the sea floor between these continents had been created by volcanism along a narrow ridge (which indeed it was). By 1972, it was believed that mid-ocean ridges are the sites of 90% of Earth's volcanic activity, and spreading rates were found to range typically between about 2 and 18 cm yr^{-1} . A further observation was that of a remarkable transition in ridge morphology at the intermediate spreading rate of around 6 cm yr^{-1} : the 'median valley' (2–3 km deep, 20–30 km broad) typical of slower rates changes to a structure some 200 m high and 2 km broad. The ocean crust created by volcanism at all well-studied ridges was found to be a fairly uniform 6.5 km thick^{2,3}, except near volcanic hotspots such as Iceland, where it could be twice that value.

Also noted (but still acknowledged to be less well understood) was that ridges typically form straight, volcano-rich segments where basalt rocks are erupted and intruded between the spreading plates. These segments are oriented perpendicular to the direction of plate spreading, and are offset by features known as transform faults, where the two plates slip sideways past each other (the San Andreas Fault is the most familiar example of a transform fault, albeit in an atypical continental context). In the ocean basins, transform faults typically lie within a rugged 'fracture-zone valley' that is strikingly similar in width and relief to the median valley seen at slower-spreading ridges.

This overall picture was refined in the 1980s and 1990s, leading to better quantification of mid-ocean-ridge dynamics. But many of these studies focused on details of ridge segmentation, the magmatic plumbing

of ridge volcanic centres, and the hydrothermal activity and associated organisms that inhabit these active deep-sea sites. Little attention was paid to the slowest-spreading ridges, largely because they are especially tough to study. For instance, the Gakkel ridge lies beneath the Arctic Ocean, investigations in which often require the assistance of ice-breakers. And the southwest Indian ridge underlies the southern oceans, halfway between southern Africa and Antarctica, which experience some of the most extreme weather in the world.

These logistical issues also mean a higher cost for field research, which makes it more difficult to justify research proposals within the modern peer-review funding process. Geological 'fishing' expeditions went out of fashion in the late 1970s in favour of hypothesis-driven science in more accessible regions. So even though there was evidence in the early 1980s that the ocean volcanic crust appeared to be much thinner than normal at an ultraslow-spreading ridge⁴, this intriguing observation was largely neglected. As a recent burst of publications^{5–7} attests, however, ultraslow ridges are at last being subjected to modern marine field research. The paper by Dick *et al.*¹ is the latest in line, and documents striking contrasts between such ridges and their faster-moving cousins.

There are two especially notable differences. One is the apparent absence of volcanic activity on long stretches of ridge (for example, there is no apparent volcanism along an 80-km section of the Gakkel ridge). The other is the frequent occurrence of oblique, 'amagmatic' segments instead of transform faults. The existence of these segments, which are nonetheless producing ocean floor and spreading, is a challenge to current theories of how rock melts in the Earth's mantle and migrates beneath ridges to create observed patterns of volcanism.

If melting has indeed ceased beneath these segments of the ridge, there are three possible explanations. First, that melting is taking place at depths of less than about 25 km, which can be cooled by heat conduction to the surface at ultraslow upwelling rates, instead of the 25–60-km melting layer inferred from the chemistry of mid-ocean-ridge basalts. Second, that the mantle temperature beneath amagmatic segments is anomalously cool, and so fails to produce melts (an idea proposed to explain the

occasional absence of melting at so-called 'non-volcanic rifted margins'⁸). Third, that the temperature at which this particular mantle material would melt is anomalously high. All of these possibilities seem strange. But they are testable.

The last two would also imply that there is correlation between lack of melting and ultraslow spreading, and that ridges that don't melt have a greater viscous resistance to spreading apart — an idea in apparent conflict with the well-accepted concept that ridges are a passive response to the separation of tectonic plates. Alternatively, the absence of volcanism could be a by-product of the lateral migration of melts in the mantle beneath the ridge, for at least 40 km. If this is so, then the mantle rocks collected along these segments should show evidence that they underwent the same amount of melting as mantle rocks dredged from faster-spreading ridges. Erupted basalts, where found, should also look similar to basalts erupted at faster-spreading ridges — and they do not, as Dick *et al.*¹ discuss.

What about the presence of oblique, amagmatic spreading segments at ultraslow ridges? The contrast with transform faults could be a telling observation. It may finally allow us to figure out why transform faults are so common at ridges, and why they are almost invariably found within a rugged fracture-zone valley. The best partial answer I've heard to the question 'Why fracture zones?' is a suggestion⁹ put forward almost 30 years ago — that such zones form in response to contraction, parallel to the ridge, induced by the cooling of the new plate as it ages and moves away from a ridge. At ultraslow, oblique spreading segments, perhaps sea water can penetrate along faults to a much greater depth than at other, faster-spreading ridges? The consequent reaction between water and mantle rocks would transform high-density mantle into serpentine, a rock type of much lower density, that could 'fill' the component of plate contraction along the ridge axis.

Answers to the geological puzzles raised by ultraslow ridges are still unclear. But it's a sure bet that such ridges will be getting a lot more attention over the next few years. ■

Jason Phipps Morgan is in the Department of Marine Geodynamics, GEOMAR, Wischhofstrasse 1-3, Kiel D-24148, Germany. e-mail: jmorgan@geomar.de

1. Dick, H. J. B., Lin, J. & Schouten, H. *Nature* **426**, 405–412 (2003).
2. Chen, Y. J. *Geophys. Res. Lett.* **19**, 753–756 (1992).
3. White, R. S., McKenzie, D. P. & O'Nions, K. J. *Geophys. Res.* **97**, 19683–19715 (1992).
4. Reid, I. & Jackson, H. R. *Mar. Geophys. Res.* **5**, 165–172 (1981).
5. Michael, P. J. *et al.* *Nature* **423**, 956–961 (2003).
6. Jokat, W. *et al.* *Nature* **423**, 962–965 (2003).
7. Edmonds, H. N. *et al.* *Nature* **421**, 252–256 (2003).
8. Reston, T. J. & Phipps Morgan, J. *Geology* **32**, 133–136 (2004).
9. Turcotte, D. L. *J. Geophys. Res.* **79**, 2573–2577 (1974).

Optics

X-ray microscopy lacks focus

Phys. Rev. Lett. **91**, 204801 (2003)

Because diffraction effects generally limit the resolution of a microscopic technique to roughly the same size as the wavelength of the radiation used, an 'optical' microscope with atomic resolution needs to use X-rays. But C. Bergemann and colleagues show that, even for X-rays with wavelengths of about the same size as atoms (0.1 nm or so), a beam can't be focused down to a spot smaller than around 10 nm.

Although it is challenging to make lenses and other optical components for X-rays, beams just 40 nm across have so far been produced. Bergemann *et al.* point out that focusing X-rays in a narrow waveguide is equivalent to confining a quantum wavefunction in a box, for which the uncertainty principle sets a fundamental limit on the degree of localization. In effect, a focusing waveguide such as a tapering wedge becomes leaky once its width falls below a critical size of at least 10 nm. So the beam widths of 1 nm or so desired for techniques such as single-atom X-ray fluorescence are ruled out.

Philip Ball

Immunology

Long gone, but not forgotten

J. Immunol. **171**, 4969–4973 (2003)

Immune cells remember a smallpox vaccination even after decades, according to Shane Crotty and colleagues. Like first love, the encounter is engraved in the memory for ever.

Most antiviral vaccines elicit the production of plasma cells and memory B cells, and both offer long-term protection;

plasma cells sustain serum levels of specific antibodies, and memory B cells drive a rapid antibody response after a new encounter with the pathogen. The smallpox vaccine was very effective and resulted in eradication of the disease worldwide. But does memory last when the virus is long gone?

Crotty *et al.* found smallpox-specific antibodies and memory B cells in the blood of individuals who were last vaccinated between 1 and 60 years previously. The number of memory B cells declined shortly after vaccination, but then reached a plateau and remained steady for 50-plus years. Following a booster vaccination, the memory cells produced a powerful antibody response — even after decades of quiescence. The T cells that support the B cells had also remembered the brief encounter from the past.

In view of possible bioterrorism threats, this long-lived immunological memory may provide some comfort. Still, it has not been shown (and experimentally cannot be) that vaccinated individuals are protected against a smallpox attack.

Marie-Thérèse Heemels

Atmospheric physics

Breaking waves

Geophys. Res. Lett. doi:10.1029/2003GL018207 (2003)

There are gravity waves and gravity waves. One type is the elusive evidence for the echoes of cosmological events such as collisions between black holes. Another is more familiar, at least in its visible consequences of creating bands of clouds. These are atmospheric gravity waves — oscillating parcels of air, teetering between buoyant and gravitational forces, and created, for example, when wind sweeps across a mountain range.

For studying the behaviour of this form of gravity wave high in the atmosphere, James A. Whiteway and colleagues had the mountains (near Aberystwyth in Wales) and ground-based instrumentation. They also had the services of a research aircraft, the Egrett, equipped for measuring atmospheric turbulence at high resolution. The plane can fly very slowly and very high, into the lower stratosphere.

Whiteway and colleagues' results reveal a gravity wave breaking, in a process called overturning, at a height of 11–12 km. The wider significance of such measurements is how wave-breaking might affect conditions at the tropopause, where the troposphere meets the stratosphere. On the evidence of the group's data on temperature, water vapour and ozone, the waves do produce substantial mixing in this part of the atmosphere.

Tim Lincoln

Molecular biology

SUMO wrestles with genes

Proc. Natl Acad. Sci. USA **100**, 13225–13230 (2003)

In the cell nucleus, DNA is packaged into a structure that somewhat resembles a string of beads, each bead being made of DNA wrapped around a core of histone proteins. Histones can be modified — enzymes can add, remove or alter chemical groups — to change gene function. Now Yuzuru Shiio and Robert N. Eisenman describe an additional family of proteins that can alter histones and may switch genes off.

Shiio and Eisenman find that proteins from the small ubiquitin-related modifier (SUMO) family bind to histones *in vivo* and *in vitro*. The molecules are thought to attach to histone tails, exposed amino acids that protrude from the core. SUMO binding correlates with transcriptional repression — less DNA is copied into messenger RNA — so SUMO may act to silence genes.

Structurally, SUMO resembles the omnipresent ubiquitin, which marks proteins out to be destroyed. When ubiquitin binds to histones, genes can be turned on. Ubiquitin and SUMO may act together, like a counterbalance, to regulate gene function.

Helen R. Pilcher

Polymer chemistry

Plastic wrap for nanocircuits

J. Am. Chem. Soc. **125**, 14113–14119 (2003)

As microelectronics gets smaller, the demands on insulating materials get ever tougher. At the nanometre scales envisaged for circuit components in the coming decade, the standard insulator of silicon technology — silicon dioxide — will become leaky, giving rise to crosstalk and device breakdown. Various alternative materials are being explored that have better insulating properties, providing lower dielectric constants.

A class of candidates now reported by Timothy M. Long and Timothy M. Swager has several attractions. They are organic polymers, processible from solution rather than requiring the high-vacuum methods of inorganic materials. They resist water absorption (water greatly increases the dielectric constant), are potentially cheap and robust, and won't break down at high temperatures. And their properties can be fine-tuned by chemical design.

The polymers contain triptycene molecular units attached to a carbon backbone. These bulky, rigid, propeller-shaped groups prevent the polymer chains from packing tightly, and the resulting 'free volume' creates dielectric constants low enough to cope with circuit features at least as small as 100 nanometres.

Philip Ball



Lasting memories: vaccination against smallpox in 1939.

Uncoupling the agony from ecstasy

A mitochondrial protein may mediate a dangerous side-effect of some recreational drugs.

The recreational use of amphetamine-type stimulants can produce a marked and sometimes lethal increase in body temperature. Here we show that mice deficient in a mitochondrial protein known as UCP-3 (for 'uncoupling protein-3') have a diminished thermogenic response to the drug MDMA (3,4-methylenedioxyamphetamine, nicknamed 'ecstasy') and so are protected against this dangerously toxic effect. Our findings indicate that UCP-3 is important in MDMA-induced hyperthermia and point to a new therapeutic direction for solving an increasing public-health problem.

A United Nations report estimates that there has been a 70% increase in the worldwide recreational use of MDMA (Fig. 1) from 1995 to 2000, with a consequent increase in the number of hospital cases and deaths^{1,2}. Most deaths result from a syndrome of persistent hyperthermia which leads to the breakdown of skeletal muscle (rhabdomyolysis), with subsequent kidney and other organ-system failure. The molecular basis for MDMA-induced hyperthermia is unknown.

Skeletal muscle may have a role in adaptive thermogenesis³, with heat being generated through ryanodine-receptor-mediated calcium cycling and, possibly, as a result of the action of mitochondrial uncoupling proteins. These proteins divert the energy produced by cellular respiration from driving ATP synthesis into producing heat instead. Brown fat, for example, is rich in mitochondria containing uncoupling protein-1 (UCP-1), which is an important thermoregulatory mediator.

Although the related proteins UCP-2 and UCP-3 share more than 50% sequence identity with UCP-1, it is unclear whether they also participate in thermoregulation³.



Figure 1 Hot stuff: tablets of ecstasy, popular among clubbers, can trigger a potentially lethal increase in users' body temperature.

For example, mice bred to lack UCP-3 (UCP-3^{-/-}, or UCP-3 'knockout' mice) have normal basal body temperatures and adapt appropriately to cold exposure^{4,5}, suggesting that the function of UCP-3 is not to regulate body temperature under normal physiological conditions. We therefore investigated whether this uncoupling protein could mediate the pathological thermogenic response that occurs as a result of MDMA toxicity.

Because UCP-3 is predominantly expressed in skeletal muscle and MDMA markedly increases the temperature of skeletal muscle⁶, we investigated the thermogenic response to this agent in wild-type compared with UCP-3^{-/-} mice. We found that wild-type mice showed a dose-dependent increase in skeletal-muscle temperature after administration of MDMA (Fig. 2a): their temperature peaked within the first hour after administration but remained raised for at least two hours; their rectal temperatures

also showed a similar response (Fig. 2b).

In contrast, although the baseline temperature of UCP-3^{-/-} mice was indistinguishable from that of wild-type animals (wild-type mice, 38.6 ± 0.08 °C; UCP-3^{-/-} mice, 38.7 ± 0.08 °C; *P*, not significant), these mice showed a significantly smaller rise in both skeletal-muscle and rectal temperature following MDMA administration (Fig. 2a,b).

As the skeletal muscle of UCP-3^{-/-} mice still expresses small amounts of UCP-2 (refs 4,5), UCP-2 may be mediating the slight temperature increase seen in the UCP-3^{-/-} mice. Alternatively, other thermogenic mechanisms unrelated to uncoupling (such as vasoconstriction mediated by the α₁-adrenergic receptor, for example) may be responsible⁶.

Given that the fatal effects of MDMA in both rodents and humans correlate with the degree of hyperthermia⁷, we tested whether UCP-3^{-/-} animals were protected against a

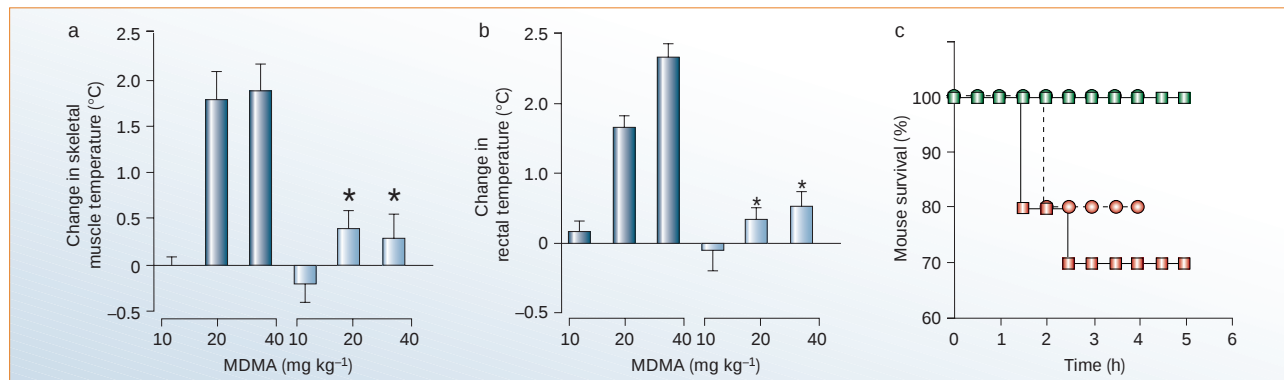


Figure 2 Mice deficient in the mitochondrial uncoupling protein UCP-3 (UCP-3^{-/-} mice) are protected against the hyperthermic effects of the drug MDMA. **a, b**, Group-housed wild-type (dark bars) and UCP-3^{-/-} (light bars) mice were injected with increasing doses of MDMA and the temperature of their skeletal muscle (**a**) and rectum (**b**) recorded 1 h later. Temperatures are plotted as the change from baseline temperatures before MDMA administration. Asterisks indicate that *P* < 0.01 when compared with wild-type controls. **c**, Percentage of mice surviving MDMA administration (circles, 20 mg kg⁻¹; squares, 40 mg kg⁻¹): red symbols, wild-type mice; green symbols, UCP-3^{-/-} mice. UCP-3^{-/-} mice were provided by M. Reitman. Further experimental details are available from the authors.

lethal challenge. Although the toxicity of MDMA in mice is less consistent than in other species⁷, a 25% lethal dose (LD₂₅) is estimated as about 50 mg of drug per kg body weight⁸. Consistent with this, wild-type animals given increasing doses of MDMA under defined conditions succumbed within 4 h of drug administration, with one in five of the animals dying at a dose of 20 mg kg⁻¹ and three in ten dying at 40 mg kg⁻¹ (Fig. 2c). However, none of the UCP-3^{-/-} animals died after administration of either dose ($n = 5$ and $n = 10$, respectively).

Our results indicate that UCP-3 is required for the rise in skeletal and core temperature that is associated with MDMA administration. Endogenously produced reactive aldehydes can also stimulate UCP-3 activity⁹, and there are some structural similarities between this class of molecule and MDMA and/or its metabolites⁷. Further investigation is needed to determine whether ecstasy can directly stimulate uncoupling activity.

The temperature response in individuals who take excessive amounts of MDMA varies, and this variation could relate to the uncoupling activity of their skeletal muscle. For example, it is known that thyroid-hormone treatment can exacerbate the hyperthermic effects of amphetamines¹⁰ and that UCP-3 is upregulated by thyroid hormone¹¹. It is possible that UCP-2 and/or UCP-3 could also mediate the toxicity of agents such as ephedrine, methamphetamine and cocaine, recreational stimulants that can produce a similar hyperthermic response and which are related to MDMA. Our demonstration that UCP-3 is a molecular mediator of the thermogenic response to ecstasy suggests that agents designed to target uncoupling proteins could provide the basis for an important new therapeutic direction.

Edward M. Mills*, Matthew L. Banks†, Jon E. Sprague†, Toren Finkel*

*Cardiovascular Branch, NHLBI, National Institutes of Health, Bethesda, Maryland 20892-1622, USA

†Department of Pharmaceutical and Biomedical Sciences, Ohio Northern University, Ada, Ohio 45810, USA

e-mail: j-sprague@onu.edu

- Landry, M. J. *J. Psychopharmacol.* **34**, 163–169 (2002).
- Dar, K. J. & McBrien, M. E. *Intensive Care Med.* **22**, 995–996 (1996).
- Lowell, B. B. & Spiegelman, B. M. *Nature* **404**, 652–660 (2000).
- Gong, D.-W. et al. *J. Biol. Chem.* **275**, 16251–16257 (2000).
- Vidal-Puig, A. et al. *J. Biol. Chem.* **275**, 16258–16266 (2000).
- Sprague, J. E., Banks, M. L., Cook, V. J. & Mills, E. M. *J. Pharm. Exp. Therap.* **305**, 159–166 (2003).
- Green, A. R. et al. *Pharm. Rev.* **55**, 463–508 (2003).
- Fantegrossi, W. E. et al. *Psychopharmacology* **166**, 202–211 (2003).
- Echta, K. S. et al. *EMBO J.* **16**, 4103–4110 (2003).
- Halpern, B. N., Drudi-Baracco, C. & Bessirard, D. *Nature* **204**, 387–388 (1964).
- Gong, D. W., Karas, M. & Reitman, M. J. *Biol. Chem.* **272**, 24119–24132 (1997).

Competing financial interests: declared none.

Photonic crystals

Imaging by flat lens using negative refraction

The positive refractive index of conventional optical lenses means that they need curved surfaces to form an image, whereas a negative index of refraction allows a flat slab of a material to behave as a lens and focus electromagnetic waves to produce a real image¹. Here we demonstrate this unique feature of imaging by a flat lens, using the phenomenon of negative refraction in a photonic crystalline material. The key advance that enabled us to make this observation lies in the design of a photonic crystal^{2,3} with suitable dispersion characteristics to achieve negative refraction over a wide range of angles.

Although negative refraction has been demonstrated at microwave frequencies in a quasi-homogeneous metamaterial⁴, imaging by a flat lens is severely constrained because of large dissipation^{5,6} and anisotropy in the metamaterial. Plane-wave negative refraction at specific incident angles has been demonstrated at microwave frequencies by using a metallic photonic crystal prism⁷ and a dielectric photonic crystal⁸. However, to focus a diverging beam from a point source, the material must exhibit all-angle negative refraction² as well as low absorption.

Figure 1a shows the image of a microwave point source of frequency 9.3 GHz (wavelength, 3.22 cm) placed 2.25 cm from a two-dimensional flat lens made of a photonic crystal fabricated from an array of cylindrical alumina rods (see supplementary information). On the far side, a high-quality image is seen at a distance of 2.75 cm. Note that there is an image of similar size for the sub-wavelength source. This image is observed only in a narrow frequency range, between 9.0 and 9.4 GHz, with the best focus at 9.3 GHz. Outside this narrow band, at all other frequencies between 2 and 12 GHz, a single focus point is not seen.

These observations can be understood by examining the band structure of the photonic crystal and the corresponding equi-frequency surfaces, from which an effective refractive index, n_{eff} , can be defined and calculated⁹. In our geometry, the central axis of the diverging beam is along the (1,0) direction of the square lattice crystal. Just below the top of the second band, between 9.0 and 9.4 GHz, the equi-frequency surfaces move inwards with increasing frequency, consistent with negative n_{eff} . The condition for negative refraction, $n_{\text{eff}}(\theta) < 0$, holds for a diverging beam of sufficiently large angle, enabling the image to be formed. Figure 1 therefore also demonstrates wide-angle negative refraction by a photonic crystal.

The value of n_{eff} is necessarily angle-dependent. Inside the crystalline lens, the electromagnetic field is highly modulated and simple ray diagrams applicable in homogeneous media cannot be used. However, upon

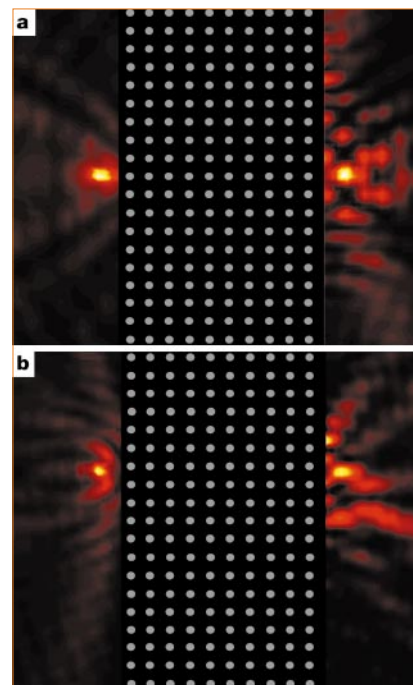


Figure 1 Imaging by a flat lens. **a**, Microwave electric-field-intensity map on a cross-sectional view of the two-dimensional source-image (left-right) system for the flat lens. **b**, Displacement of the image by 4 cm when the source is moved up by the same distance. Both panels correspond to dimensions of 37.5 × 30.0 cm. The intensity scale on the source side varies from -20 dB (yellow) to -48 dB (black), and on the image side from -30 dB to -75 dB.

emerging, the waves create a focus of the divergent beam emanating from the point object.

Conventional optical systems have a single optical axis and limited aperture, and cannot focus light onto an area smaller than a square wavelength¹. In contrast, our flat lens does not have a unique optical axis and is not restricted by aperture size. We have demonstrated both of these features by moving the source up by 4 cm: the image moves a corresponding distance in the same direction (Fig. 1b).

The unique properties of our flat lens provide new perspectives on imaging. A particular advantage of the photonic crystalline material is its scalability to submicrometre dimensions for possible applications from microwave to optical frequencies.

Patanjali V. Parimi, Wentao T. Lu, Plarenta Vodo, Srinivas Sridhar

Department of Physics and Electronic Materials Research Institute, Northeastern University, Boston, Massachusetts 02115, USA

e-mail: s.sridhar@neu.edu

- Pendry, J. B. *Phys. Rev. Lett.* **85**, 3966–3969 (2000).
- Luo, C. et al. *Phys. Rev. B* **65**, 201104(R)(4) (2002).
- Notomi, M. *Phys. Rev. B* **62**, 10696–10705 (2002).
- Shelby, R. A., Smith, D. R. & Schultz, S. *Science* **292**, 77–79 (2001).
- Houck, A. et al. *Phys. Rev. Lett.* **90**, 137401(4) (2003).
- Iyer, A. K. et al. *Opt. Express* **11**, 696–708 (2003).
- Parimi, P. V. et al. <http://arxiv.org/abs/cond-mat/0306109> (2003).
- Cubukcu, E. et al. *Nature* **423**, 604–605 (2003).
- Foteinopoulou, S. & Soukoulis, C. M. *Phys. Rev. B* **67**, 235107(5) (2003).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

An ultraslow-spreading class of ocean ridge

Henry J. B. Dick, Jian Lin & Hans Schouten

Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA

New investigations of the Southwest Indian and Arctic ridges reveal an ultraslow-spreading class of ocean ridge that is characterized by intermittent volcanism and a lack of transform faults. We find that the mantle beneath such ridges is emplaced continuously to the seafloor over large regions. The differences between ultraslow- and slow-spreading ridges are as great as those between slow- and fast-spreading ridges. The ultraslow-spreading ridges usually form at full spreading rates less than about 12 mm yr^{-1} , though their characteristics are commonly found at rates up to approximately 20 mm yr^{-1} . The ultraslow-spreading ridges consist of linked magmatic and amagmatic accretionary ridge segments. The amagmatic segments are a previously unrecognized class of accretionary plate boundary structure and can assume any orientation, with angles relative to the spreading direction ranging from orthogonal to acute. These amagmatic segments sometimes coexist with magmatic ridge segments for millions of years to form stable plate boundaries, or may displace or be displaced by transforms and magmatic ridge segments as spreading rate, mantle thermal structure and ridge geometry change.

Ocean ridges spreading at less than $\sim 20 \text{ mm yr}^{-1}$ comprise $\sim 20,000 \text{ km}$ of the $\sim 55,000 \text{ km}$ global ridge system¹, but until recently have been little studied. On the basis of studies at faster spreading rates, ocean ridges have been divided into fast-, intermediate-, and slow-spreading, each with distinctive morphologic characteristics (Fig. 1a). Slow-spreading ridges (less than $\sim 55 \text{ mm yr}^{-1}$) have deep rift valleys with highly variable relief from ~ 400 to $2,500 \text{ m}$ (ref. 2) and rough rift mountain topography weakly correlated to spreading rate³. Fast-spreading ridges (~ 80 – 180 mm yr^{-1}) have low ($\sim 400 \text{ m}$) axial highs⁴, sometimes with small linear depressions (less than $\sim 100 \text{ m}$ wide and less than $\sim 10 \text{ m}$ deep, for example⁵) at their crests, and minimal rift mountain topography uncorrelated to spreading rate. Intermediate-spreading ridges (~ 55 – 70 mm yr^{-1} ; ref. 2) have long alternating sections with either slow- or fast-spreading ridge morphology^{6–8}.

Slow-spreading ridges near mantle hotspots that have an axial rise (for example, the Reykjanes ridge), and abrupt changes in morphology at intermediate-spreading-rate ridges, show that ridge morphology is not strictly dependent on spreading rate, but may reflect variations in ridge geometry, melt supply, and mantle composition and thermal structure.

But this classification also does not take into account variations in crustal thickness with spreading rate. Although seismic crustal thickness shows little dependency on spreading rate down to 20 mm yr^{-1} , it then drops off rapidly (Fig. 1b). Models for buoyant and passive plate-driven mantle flow attribute this to conductive cooling from above depressing the top of the mantle melting column and limiting melt production^{9–12}. Thus, changes in ridge geometry, mantle composition, flow and thermal structure that at faster rates would have a minor effect, dramatically affect crustal production and tectonics at very slow spreading rates^{13,14}, creating a class of ridges that is distinct from previous categories.

Just as the East Pacific Rise is the only fast-spreading ridge, the only ultraslow-spreading ridge is the Gakkel ridge in the Arctic Ocean: an $1,800\text{-km}$ -long ridge without transform faults spreading 8 to 13 mm yr^{-1} (ref. 15). The very-slow-spreading (14 – 16 mm yr^{-1}) Southwest Indian ridge (SWIR) and the remaining Arctic ridge system (13 – 18 mm yr^{-1}) are transitional between slow and ultraslow. Just as intermediate-rate ridges alternate between fast and slow morphology, these ridges alternately exhibit slow or ultraslow ridge characteristics over long sections. Thus, there are two intermediate-rate classes rather than one, representing transitions

from fast- to slow-spreading and slow- to ultraslow-spreading tectonics (Fig. 1b).

Magmatic accretionary ridge segments are second-order segments that link together between transform faults to make up first-order segments¹⁶. They are considered the basic unit of magmatic accretion, and at slow-spreading ridges often have circular negative mantle Bouguer anomalies (MBAs) centred on them, suggesting focused melt and/or mantle flow resulting in thicker crust at their centres^{17–20}. As there are no transform faults defining a first-order segmentation at ultraslow rates, we use the term ‘super-segment’ to describe distinct physiographic regions consisting of linked magmatic and amagmatic ridge segments.

Magmatic segments form sub-perpendicularly to the least principal compressive stress, at an angle intermediate to the ridge trend and orthogonal to the spreading direction, for example^{21,22}. Morphologically, they consist of linear axial highs or troughs with saddle points at the segment centres and rift valley walls consisting of staircase normal boundary faults. In the absence of transform faults, these structures occur in sets overlapping en echelon or connect by small non-transform discontinuities to accommodate plate boundary trends that are oblique to the spreading direction. Where the ridge trend is highly oblique (for example, the Mohs ridge), intra-rift faults and axial ridges often have sigmoidal traces where they merge with rift valley bounding faults that run at higher angles sub-parallel to the ridge trend²¹. The close interaction of these two fault sets shows that they are part of an integrated system accommodating oblique extension during brittle faulting in the shallow crust (for example, ref. 22).

Amagmatic accretionary ridge segments are a key component of ultraslow-spreading ridges, replacing transform faults while extending the zone of lithospheric accretion. In contrast to magmatic segments, they assume any orientation relative to the spreading direction, and are usually marked by an axial trough rarely more than a kilometre deep that may extend 50 km or more. They have only scattered volcanics or a thin basalt carapace, virtually no seismic layer 3 (for example, ref. 23), expose abundant mantle peridotite, and have weak magnetization and a relatively positive MBA. The basic accretion unit appears to be mantle horst blocks rising up through the rift valley floor along an axis running parallel to the plate boundary trend. These are subsequently uplifted to create the axial trough walls—often creating single $\sim 14^\circ$ – 20° low-angle fault surfaces bounding lenticular peridotite ridges.

The SWIR from 9° to 25° E

The SWIR has a very oblique trend—typically $\sim 60^\circ$ from the spreading direction—and is an example of a transitional ridge between slow and ultraslow. Long sections are similar to slow-spreading ocean ridges (that is, second-order magmatic segments linked by transform faults). Other sections lack transforms and have the character of ultraslow ridges, consisting of sub-orthogonal ~ 50 -km-long magmatic segments linked by amagmatic oblique troughs (for example, ref. 24). This slow to ultraslow transitional behaviour is probably largely due to the influence of ridge geometry. As obliquity increases ridge length, mantle upwelling rate must slow proportionately (Fig. 2b, inset). This is expressed by the spreading component perpendicular to the ridge trend: the effective spreading rate for mantle upwelling (ESR). At very slow spreading rates less than $\sim 20 \text{ mm yr}^{-1}$, varying ESR (and mantle composition and temperature) creates a region in the spreading-rate spectrum where ridges can assume either a slow-spreading or an ultraslow-spreading morphology.

The two sharply contrasting supersegments along the 1,000 km of the SWIR between the Shaka fracture zone at 9° E and the Prince Edward fracture zone at 25° E , provides a case in point (Fig. 2). The eastern 'orthogonal' supersegment, mapped in ref. 25, extends 600 km from 16° to 25° E and has 330- and 191-km-long subsections oriented approximately perpendicularly to the spreading direction linked by a 150-km slightly oblique ($\sim 10^\circ$) section (Fig. 2b). From 17° E , the ridge curves smoothly to the southwest, and thus the ESR steadily decreases to the west. At 16° E there is an abrupt change in ridge morphology and lithology at an ESR of $\sim 12 \text{ mm yr}^{-1}$. Further west the ridge reaches an obliquity of 32° from the spreading direction, resulting in possibly the lowest ESR on the global ridge system (7.8 mm yr^{-1}). This continuous change in ESR without a major ridge offset provided an opportunity to examine the influence of mantle upwelling on crustal accretion at the slowest end of the spreading spectrum. Thus, in austral summer 2000, during RV *Knorr* Cruise 162, we mapped gravity, magnetics and bathymetry along the 'oblique supersegment' out to ~ 10 – 11 Myr ago or $\sim 85 \text{ km}$ from the ridge axis on a flow line, extensively sampling both supersegments.

The orthogonal supersegment is morphologically and geophysically similar to first-order ridge segments along the slow-spreading Mid-Atlantic Ridge²⁵ (MAR) (Fig. 2 legend). To complement 13 previous dredges, we made 28 more along the orthogonal supersegment (Fig. 2a): providing 26 along the axis and 15 on the rift valley walls. Dredge tracks were typically 500 m or more in length. Nine were empty, whereas 30 contained glassy pillow basalt. Gabbro and peridotite occur in only one wall dredge, consistent with the seafloor physiography, which suggested a nearly continuous volcanic carapace.

The oblique supersegment begins at the 16° E discontinuity with a 500-m drop in rift valley depth (Fig. 2a) and abrupt changes in physiography, MBA, magnetic anomaly intensity, and rocks dredged (Fig. 3). Although the spreading appears well organized, the pattern and tectonic character of the supersegment is very unlike that of 16° – 25° E , consisting of linked magmatic and amagmatic accretionary ridge segments.

The westernmost region from $9^\circ 30'$ to $11^\circ 45' \text{ E}$ is a broad area of rough topography with a short magmatic ridge segment at its intersection with the Shaka fracture zone. From there, a poorly defined rift valley and an irregular terrain run obliquely to the northeast to about the west flank of Joseph Mayes seamount. This seamount is the first of only two significant volcanic centres on the supersegment. It splits an old lineated peridotite block that projects from its northeastern and southwestern flanks on an axis perpendicular to the spreading direction, filling the rift valley to create a 2,500-m high double-peaked volcano. Five dredges on the volcano recovered fresh pillow basalt.

East of Mayes seamount, there is a 180-km-long, 4,200-m-deep oblique amagmatic segment from $11^\circ 30'$ to $14^\circ 24' \text{ E}$, oriented 32° from the spreading direction. It has an axial trough with local deep areas to 4,700 m. Twenty-two dredges (2,574 kg) recovered 46.2% serpentinized peridotites, 2.6% dunite, 14.2% pillow basalt, 19.9% basalt and diabase, and 12.6% hydrothermal silica, sepiolite, and massive sulphides from the trough walls and floor. As even a thin basalt flow effectively covers peridotite emplaced to the seafloor, the dredge statistics and lack of gabbro show that igneous crust is thin or absent here and that the terrain is largely created by faulting. Although the trough floor is rough, with ~ 500 -m relief, there is no morphologic evidence of a neovolcanic zone. Rather, long irregular axial horst blocks rise up from the valley floor. Dredging, physiography, weak magnetization, and positive MBA suggest these blocks

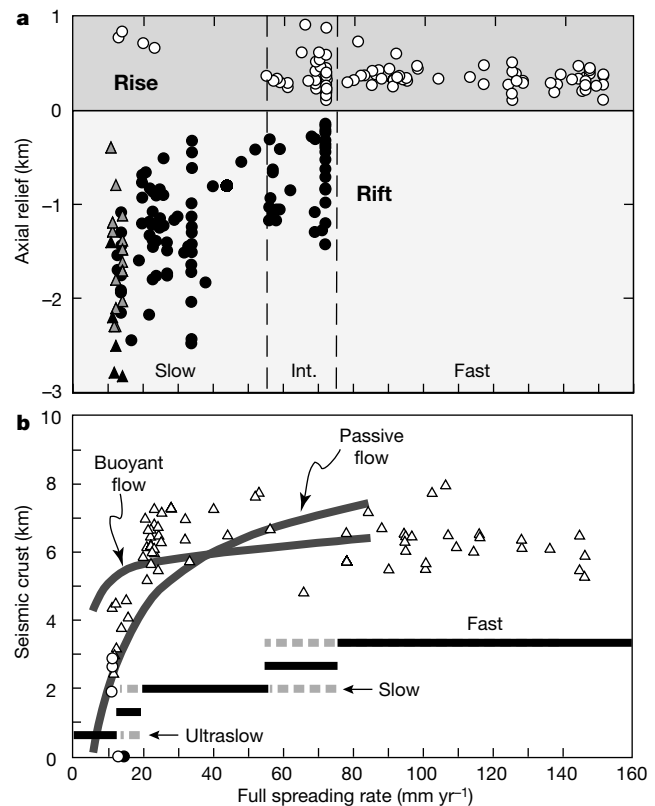


Figure 1 Ridge properties as a function of spreading rate. **a**, Axial relief versus spreading rate. Axial rises shown by white circles and axial rifts by black circles modified from ref. 2. Triangles show new Gakkel ridge¹⁵ and SWIR 9°–16° E data: grey triangles, amagmatic accretionary segments; black triangles, magmatic ridge segments. Note that the amagmatic segments have lower relief. Dashed lines show a conventional division for ocean ridges into fast-, intermediate- (Int.) and slow-spreading with intermediate-spreading ridges those consistently exhibiting alternate fast and slow behaviour. **b**, Seismic crustal thickness versus spreading rate from ref. 12 (white triangles). White circles are new Gakkel ridge values (AMORE²³); black filled circle is for SWIR 9°–16° E oblique supersegment. Zero values are based on dredging. Crustal thickness curves calculated for passive flow⁴³ and buoyant flow⁴⁹ models. A revised ridge classification scheme is shown by the heavy black straight lines giving the spreading-rate ranges for ultraslow-, slow-, fast- and the two transitional classes. The SWIR and the Knipovich ridge are examples of ultraslow- to slow-spreading transitional ridges while the intermediate-rate Southeast Indian ridge is an example of a slow- to fast-spreading transitional ridge. Transitional ridges exhibit alternating sections with the morphology of the faster- and slower-spreading classes, rather than representing intermediate characteristics between them. Broken lines indicate the typical overlap of tectonic styles for each of the three principal classes of ocean ridge that give rise to the two transitional ridge classes.

are largely partially serpentinized peridotite. The trough walls consist of a series of large irregular or lenticular blocks with smooth low-angle scarps (14° to 20°) that are highly oblique to the spreading direction, exposing abundant mantle peridotite. Many of these scarps appear to be simple fault surfaces that form continuous ramps ~ 6 km long rising 1 to 2 km from the rift valley floor to the crest of the rift mountains on which old axial horst blocks were uplifted to form linear ridges parallel to the ridge trend.

We traced hundreds of small northeastern- and southwestern-facing linear scarps and asymmetric ridges in the 50-m contour bathymetry all along the oblique supersegment, though most are too small to be visible in Fig. 3. Almost all are oriented nearly orthogonal to the spreading direction. These are interpreted as

inward- and outward-facing high-angle normal faults that are only 50–150 m high, except at the major magmatic centres. They contribute little to the trough relief, and cut the walls at an acute angle, giving them an unusual serrated appearance. These faults can run long distances across the trough floor, up the walls and into the rift mountains.

A second major volcanic segment, also oriented perpendicularly to the spreading direction, is located at $14^\circ 41' \text{ E}$ on a southeastward kink in the oblique supersegment centred on a large cross-axis topographic high. This 40-km-long ‘Narrowgate’ segment closely resembles the magmatic segments along the orthogonal supersegment with staircase normal faulting and a deep axial rift ($>2,500$ m relief). There is a well-defined axial neo-volcanic ridge, large

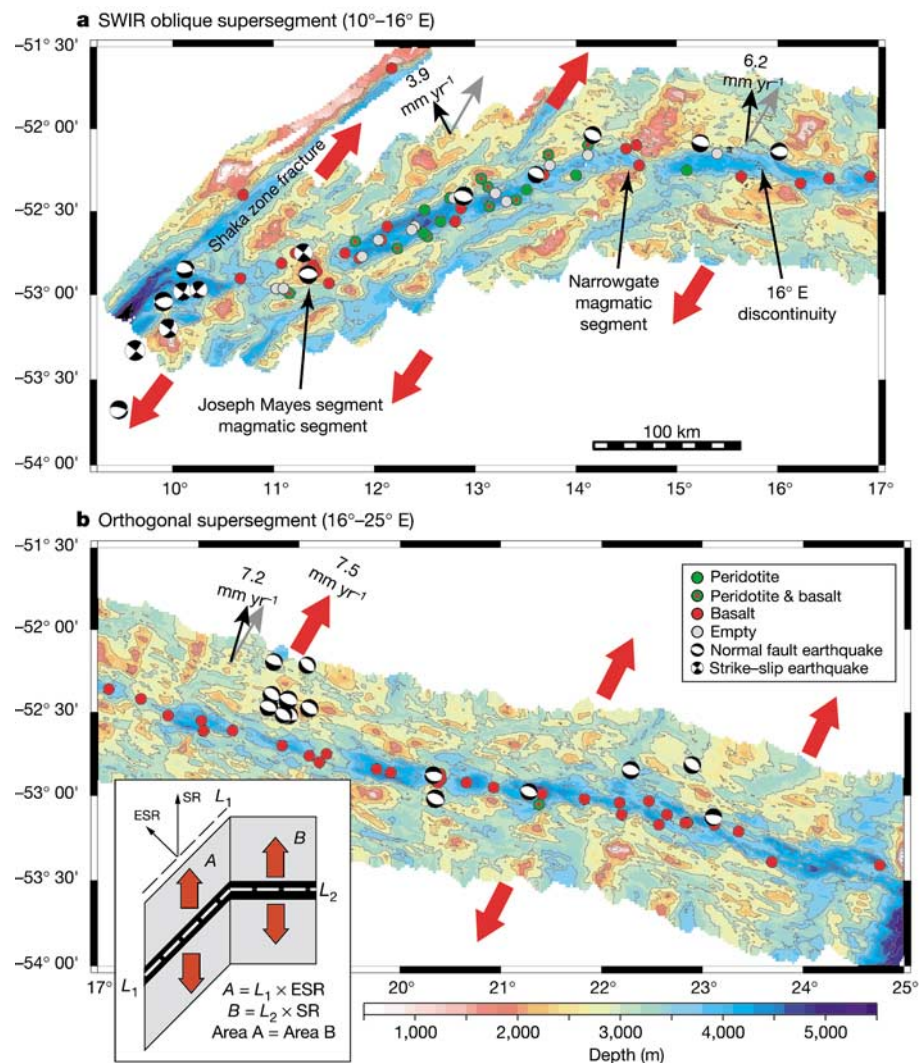


Figure 2 SWIR bathymetry from 9° to 25° E . We calculated spreading directions using fits to the SWIR transform from 2° to 37.5° E (red arrows). Fault solutions are for well-located Harvard catalogue earthquakes since 1977 of magnitude >4.5 . Small circles give dredge locations and are larger than the tracks. **a**, Oblique supersegment mapped during RV *Knorr* cruise 162. **b**, Orthogonal supersegment mapped by ref. 25. The supersegment consists of $\sim 42 \pm 15$ km en echelon second-order¹⁶ magmatic ridge segments sub-perpendicular to the spreading direction. Their average magnetization anomaly is 15 A m^{-1} (ref. 25). Distinct circular MBA lows are centred about the western ridge segments, but to the east the signal is more chaotic²⁵. This supersegment has a broader rift valley (~ 47 km) and greater cross-axis relief (~ 2.2 km), but closely resembles typical MAR spreading centres. In contrast to the oblique supersegment, rift mountain topography is broadly symmetrical north and south of the ridge, with local topographic highs and lows often matching on either side²⁵. The inset in **b** illustrates the effect of ridge

geometry on mantle upwelling. Where conductive heat loss becomes significant, the rate of mantle upwelling controls the extent of melting, lithospheric thickness and mantle rheology. Upwelling rate is a function of spreading rate (SR), ridge obliquity¹³, and type of mantle flow (for example, buoyant three-dimensional versus passive two-dimensional)⁵⁰. As a ridge becomes oblique to the spreading direction, its length increases per unit of lithosphere created, and mantle upwelling must slow proportionally to conserve mass. The ‘effective spreading rate’ for mantle upwelling (ESR), the spreading component orthogonal to the ridge trend, expresses these components. Black and grey arrows show the ESR and SR, respectively, at selected points along the ridge. Whereas the spreading rate here is nearly constant, the ESR is $\sim 14.4 \text{ mm yr}^{-1}$ at 18° E on the orthogonal supersegment, 12.4 mm yr^{-1} at the 16° E discontinuity and $\sim 7.8 \text{ mm yr}^{-1}$ at 12° E in the middle of the oblique supersegment (black arrows show half rates).

negative MBA and strong magnetization (Fig. 3). Pillow basalt was dredged from the neovolcanic zone²⁶, and basalt and diabase from the north wall.

From 14° 54' E to the first orthogonal supersegment volcano at 15° 45' E there is another oblique amagmatic accretionary segment with a deep axial trough. This trough curves gently east-southeast to connect the two supersegments. As this occurs, the ubiquitous high-

angle normal faults gradually change from cutting across the axial trough, contributing little to the terrain, to become the principal terrain-forming element. One wall dredge recovered serpentinized peridotite, while another had scoria fragments thought to be glacial erratics.

The bathymetric and lithologic discontinuities are matched by geophysical ones as well, with an abrupt increase in MBA and

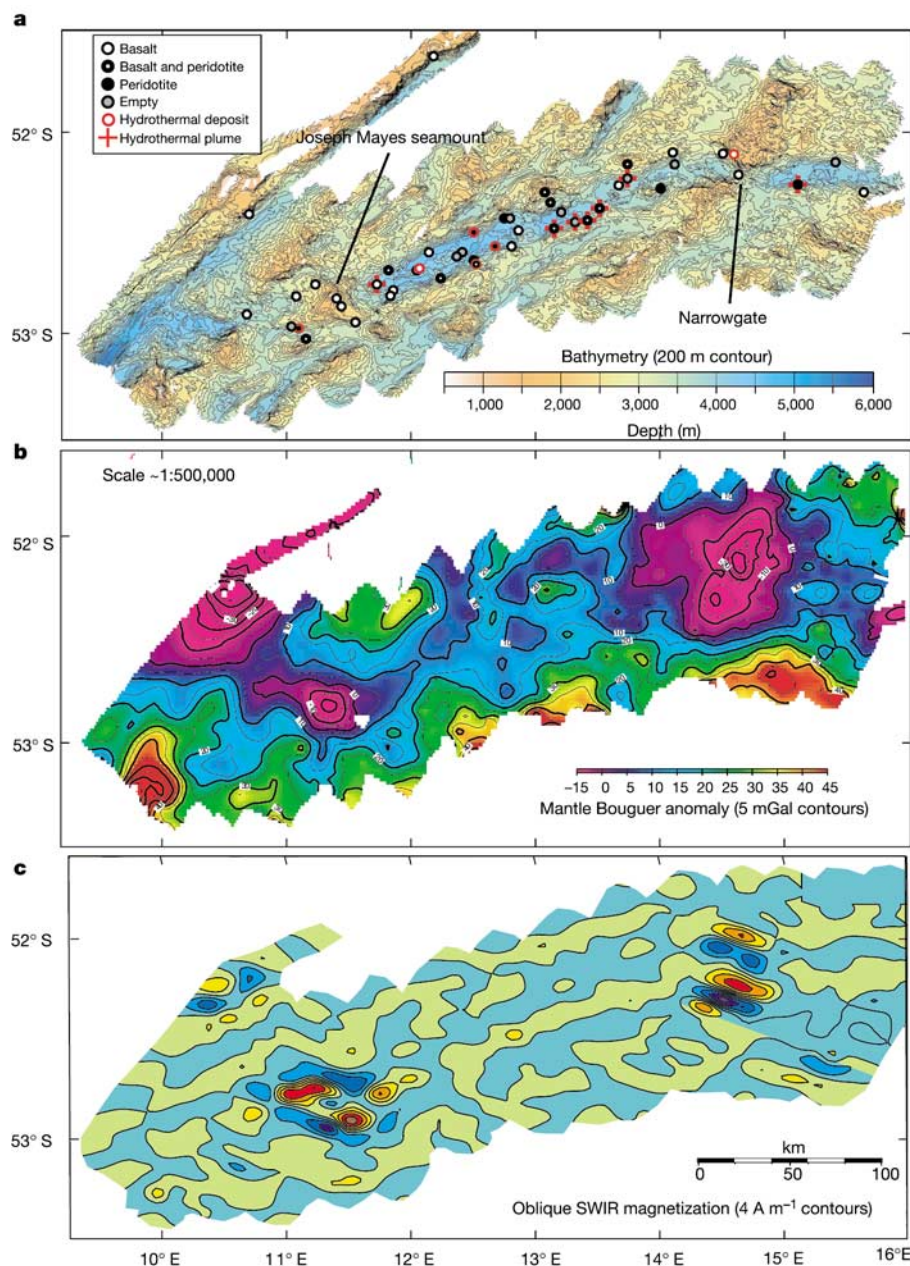


Figure 3 SWIR oblique supersegment bathymetric, mantle Bouguer and magnetization maps. **a**, 200-m contoured gridded bathymetry. **b**, Mantle Bouguer gravity anomaly. **c**, Crustal magnetization as the product of a nonlinear inversion that removed the topographic effect and the distortion resulting from the non-vertical magnetization and the Earth's magnetic field. Unlike the orthogonal supersegment, rift mountain topography is broadly asymmetric north and south of the ridge, with local topographic highs and lows rarely matching on crust of the same age on either side²⁵. A small-diameter large-magnitude MBA low is centred at Joseph Mayes seamount, which is consistent with a recently formed volcanic segment splitting an older mantle block. The other large MBA low centred at Narrowgate extends into the rift mountains along the cross-axis topographic high, consistent with a thick belt of crust formed at a long-lived magmatic segment. At the eastern end of our survey is the edge of the previously recognized large MBA low over the

first orthogonal supersegment magmatic segment²⁵ (**b**). This MBA low is similar in width but smaller in amplitude than the Narrowgate anomaly (20–25 mGal versus 35–40 mGal). 27 magnetic profiles were collected across the oblique supersegment out to about anomaly 5 (10 Myr ago). Only a few have anomalies resembling forward models of a symmetric geomagnetic reversal sequence, mostly centred near the Narrowgate cross-axis volcanic high. The best estimates of spreading rate are from lines over the Narrowgate segment and just east of the Shaka fracture zone. These give symmetric anomaly 2A full spreading rates of 14 mm yr⁻¹ and 15 mm yr⁻¹, respectively. Note that the unusual sigmoidal shape of Joseph Mayes seamount appears to be due to fissure eruptions from the peaks that created volcanic ridges extending sub-perpendicular to the spreading direction to the west and east respectively.

decrease in crustal magnetization west of 16° E. West of 16° E there are only two prominent circular MBA lows, each centred over a volcanic centre. The low over Narrowgate extends out over the northeast-southwest cross-axis topographic high, consistent with a belt of thick crust (Fig. 3b). The MBA lows correspond to crustal magnetization highs similar to those at MAR magmatic ridge segments, with magnetization ranging from -15 to $+15 \text{ A m}^{-1}$ (Fig. 3c). By contrast, the oblique axial troughs have high MBAs, and low magnetization with weak linear axial anomalies, indicating thin or missing crust. Teleseismic earthquakes give orthogonal normal fault solutions, except for a strike-slip solution at Mayes seamount that may be related to rotations accompanying splitting of the peridotite block (Fig. 2). The ubiquitous high-angle normal faults are probable sources for the remaining earthquakes.

The oblique supersegment rift mountains, unlike the orthogonal supersegment, are highly asymmetric across the rift and very rough. They have regions of lenticular blocks elongated parallel to the ridge trend and oblique to the spreading direction, as well as contrasting regions with elevated, irregularly shaped plateaus and seamounts with a lineated fabric sub-perpendicular to the spreading direction. The lenticular blocks have moderate to strongly positive MBAs similar to those of the mantle blocks on the axial trough walls. The plateaus and seamounts have weak negative MBA and morphologically appear volcanic, dominating the northern and large portions of the southern rift mountains (Fig. 3b). There is also an abandoned

fracture zone trace at 13.7° E. This and eastward V-shaped topographic trends in the rift mountains centred on Narrowgate and the 16° E magmatic segment suggest that the oblique amagmatic accretionary segments formed by the eastward migration of these magmatic ridge segments, and the gradual disappearance of others. Before ~ 8 Myr ago, then, this section of the SWIR seems to have consisted of magmatic ridge segments linked by transforms.

The Arctic ridge system

The Arctic ridge system (Fig. 4) from Iceland to 3° E on the Gakkel ridge is a transitional ridge like the SWIR, with spreading rate decreasing northward from 18 mm yr^{-1} to 12.7 mm yr^{-1} (refs 27, 28). It consists of long sections, each with their own name and character, that alternate between ultraslow- and slow-spreading physiography. In particular, the Knipovitch ridge is a ~ 550 -km-long supersegment trending 41° to 55° from the spreading direction. The average ESRs of 11.2 and 12.3 mm yr^{-1} in the north and south, respectively, fall at or below the ultraslow-spreading ESR threshold at 16° E on the SWIR. This supersegment contains no transforms and has all the physiographic features of the SWIR 10°–16° E oblique supersegment, including point-source volcanic centres and cross-axis bathymetric highs with MBA lows oriented perpendicularly to the spreading direction, linked by long deep oblique segments²⁹. Heavy sedimentation, however, obscures many features, particularly between the volcanic centres, and limited dredging largely to the highs, which, like those at the SWIR oblique supersegment, yielded largely basalt.

Gakkel ridge, although it is the slowest-spreading ocean ridge, has no transform faults anywhere along its length^{15,30}. Its spreading rate varies smoothly eastward as it nears its Euler pole from 12.7 mm yr^{-1} to 6.0 mm yr^{-1} . With a near-orthogonal trend, the ESR equals the absolute spreading rate along most of it. The Gakkel ridge's trend conforms to the shape of the conjugate Lomonosov ridge and the Barents shelf margins, so, unlike the SWIR, its geometry is inherited from early continental rifting^{15,31}. The 220-km-long faster section west of 3° E, like the SWIR from 16° to 25° E, consists of a string of magmatic ridge segments³² lying in a deep rift valley bounded by high-angle normal faults. Abundant pillow basalt was dredged there³², though relatively small magnetic anomalies and seismic measurements indicate a thin crust (~ 2 – 3 km)^{10,33–35}. At 3° E, at a small non-transform offset, the rift valley floor drops 1 km, and only peridotite was dredged for 80 km to the east^{32,36}. Scattered volcanics and peridotites then dominate to 13° E, where the first of a series of isolated magmatic segments occurs. These increase in number to 72° E, with mantle peridotite occurring rarely east of 40° E (ref. 32). East of 72° E volcanism may again diminish, as the rift is heavily sedimented and only two additional magmatic segments have been identified at 84° E and 92° E (ref. 37), while gravity measurements at 95° E suggest the crust is vanishingly thin there³⁸.

Bathymetry^{15,23,32}, magnetics^{23,31,39}, and sidescans^{15,37} confirm the highly localized volcanism east of 3° E. This takes three forms: (1) volcanic segments centred on large cross-axis ridge-perpendicular highs, (2) large point-source volcanoes, and (3) large highly reflective flows locally ponded on the rift valley floor near major volcanic centres¹⁵. The ponded flows are probably thin, given their low magnetization^{23,39}. Similar to the SWIR Narrowgate segment, the magnetization increases sharply at cross-axis volcanic highs and local axial highs³¹, and rift valley wall morphology indicates large tectonic uplifts on high-angle normal faults. The rift valley floors shoal $\sim 2 \text{ km}$ to the midpoints of the magmatic segments, which are centred on prominent local MBA lows.

Between the volcanic centres are long deep rift valleys ($\sim 4,600$ – $5,400 \text{ m}$) with weakly negative to slightly positive MBA¹⁵ and low magnetic intensity^{13,31}. Here the rift walls consist of long linear ridges or more irregular massive block uplifts exposing abundant peridotites^{15,36} identical to faults bounding the 9°–16° E SWIR

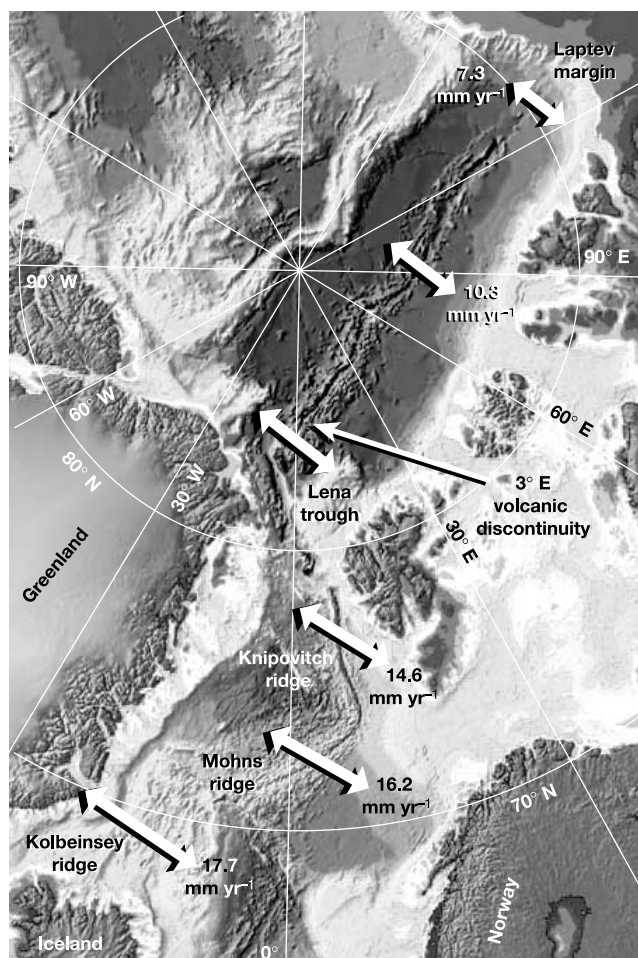


Figure 4 Arctic ridge system modified from the IBCAO bathymetry⁵¹ showing its subdivisions including the Gakkel ridge and the Knipovitch ridge (see text). Also shown for reference is the position of the 3° E volcanic discontinuity. Spreading rates and azimuths are calculated from the Nuvel 1 model²⁷.

amagmatic segments. They are 10–15 km across and 1–2 km high, with single low-angle scarps dipping 10° to 20° from the crest of the rift mountains to the valley floor^{15,32}. Although most of the magmatic and amagmatic segments are orthogonal to the spreading direction, there is a 350-km section between 35° and 77° E where the ridge bends sharply to the north and back again, before resuming an orthogonal trend. Around the bend, magmatic segments remain oriented perpendicularly to the spreading direction, while amagmatic segments follow the ridge trend as at the SWIR oblique supersegment.

Discussion

The traditional division of ocean ridges is inadequate to characterize their full variability. Examples of ridges spreading less than ~20 mm yr⁻¹ show a common set of unusual tectonic features sufficient to characterize a new class of spreading ridges. Ultraslow-spreading ridges consist of linked magmatic and amagmatic accretionary segments, lack transform faults, and have discontinuous volcanism with large regions of highly attenuated or missing crust. Where the ESR varies continuously at the Gakkel ridge and the

SWIR there are discontinuities at 12.4 and 12.0 mm yr⁻¹ where magmatism stops abruptly, and the ridges shift from slow-spreading to ultraslow-spreading behaviour. Elsewhere on the SWIR and Arctic ridge system, we find ultraslow-spreading characteristics appearing when the ESR is less than ~12 mm yr⁻¹. Thus, we pick 12 mm yr⁻¹ absolute spreading rate as the usual boundary for ultraslow-spreading ridges, noting that changes in ridge geometry and mantle temperature and composition can cause ultraslow-spreading tectonics up to about 20 mm yr⁻¹.

A geochemical discontinuity matches the physiographic and crustal changes as well. Axial basalts on the faster-spreading side are generally typical MORB, and have lower Na₈ and higher Fe₈, while those on the slower side may be alkaline and/or isotopically enriched with high LREE and other incompatible trace-element abundances^{26,32,40}. Na₈ and Fe₈ are the weight% Na₂O and FeO corrected for fractionation to 8 wt% MgO. High Na₈ may reflect large degrees of mantle melting, whereas high Fe₈ may reflect deep mantle melting⁴¹. This is consistent with dramatically lower extents of mantle melting beneath ultraslow ridges. Simple conductive models for variations in mantle melting with spreading rate do not predict a volcanic discontinuity of ~12 mm yr⁻¹ (refs 9, 11). This implies a threshold behaviour in mantle dynamics and melt generation, perhaps representing a switch from regions of continuous volcanism dominated by buoyant flow to regions dominated by passive flow.

Volcanism along ultraslow ridges, however, does not correlate with ESR. Long-lived volcanic centres tend to occur at bends at both the Gakkel ridge and the SWIR, suggesting a physical control on where melting occurs. The spatial distribution of mantle chemical heterogeneities also probably affects volcanism. Mayes seamount, for example, lies some 300 km west of the SWIR 16° E discontinuity and is possibly the largest non-hotspot-related axial volcano on the global ridge system. This may be explained by melting of a heterogeneous veined mantle. The isotopic composition of basalts and peridotites from the SWIR oblique supersegment are consistent with low degrees of melting of a veined mantle⁴², while basalts erupted east of the 16° E discontinuity on the orthogonal supersegment are consistent with higher-degree melts dominated by a depleted host mantle peridotite⁴³. Pyroxenite with a mid-ocean-ridge basalt (MORB)-like composition melts 35 to 50 km deeper than peridotite, and can be 60% molten at the peridotite solidus⁴⁴. Thus when thermal conduction is sufficient to inhibit significant peridotite melting, there would still be vein melts that could aggregate and trigger local mantle diapirism. Excess enriched vein-melts aggregated from a broader region than usual for ridge volcanism could explain the enriched isotopic and trace-element composition of many basalts erupted at ultraslow ridges. Volcanism, then, would largely be a function of regional mantle composition and temperature, rather than spreading rate.

The abrupt shift to normal fault dominated terrain and the reorientation of the axis of faulting to perpendicular to the spreading direction at the magmatic segments at oblique ultraslow-spreading ridges is consistent with plate failure in the brittle regime, and suggests a major change in mantle dynamics. Despite their atypical chemistry, ultraslow-spreading volcanic centres are often as large as, or larger than, those at slow-spreading ridges. This heightened melt production suggests that mantle flow beneath these magmatic centres may have a buoyant component, whereas there is largely plate-driven passive flow beneath the amagmatic segments. Buoyant flow dramatically reduces the effects of conductive heat loss and would produce a thinner weaker lithosphere. More melt would also weaken the shallow lithosphere, notably by injection of magma in dykes, forming weak fluid-filled cracks. Gabbroic dykes found in East Pacific Rise mantle at Hess Deep⁴⁵ show that such cracks could extend at least 6 or 7 km in depth.

Amagmatic segments are identified here as a previously unrecognized class of accretionary plate boundary structure. On the

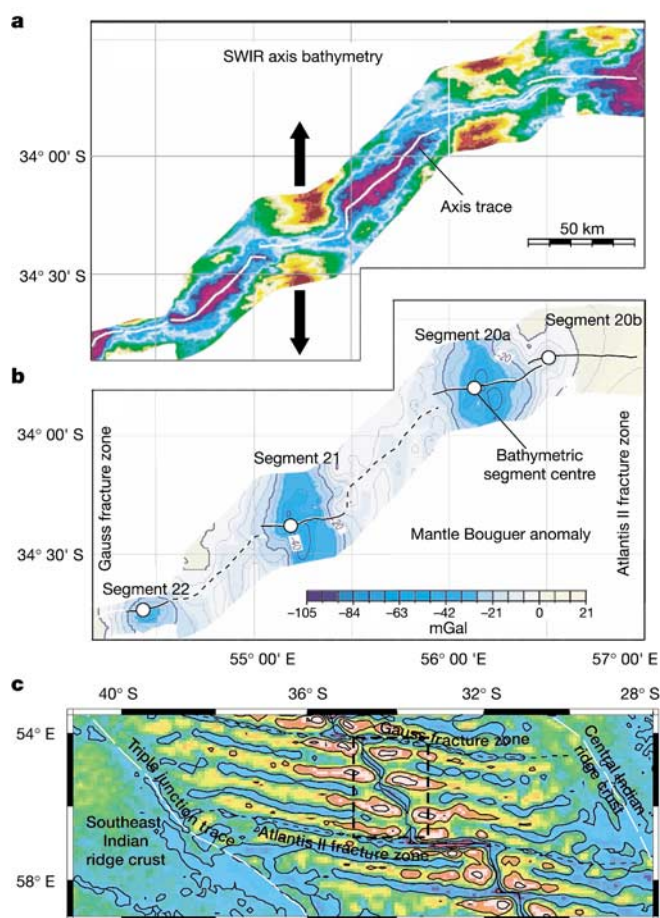


Figure 5 SWIR axial region between the Atlantis II and Gauss fracture zones (modified from ref. 24). **a**, Bathymetric map of the ridge axis showing greatest depths in blues and purples and shallowest in reds and yellows. Total depth range is from 6,500 to 1,850 m. **b**, Mantle Bouguer gravity anomaly map with 5-mGal contours. Trace of magmatic segments shown as thin black lines; trace of oblique amagmatic accretionary segments shown as dotted lines. **c**, Satellite free-air gravity map showing the location of the map areas in **a** and **b** (dashed box) and the traces of the oblique amagmatic accretionary segments across the Indian Ocean floor extending to the trace of the Rodriguez triple junction. These traces (thick dashed white line) separate the SWIR from the central and Southeast Indian ridge crust that formed as the triple junction migrated to the northeast with time. Spreading directions run parallel to the adjacent Atlantis II and Gauss transforms as indicated by the thick dotted lines.

Gakkel ridge most amagmatic segments are orthogonal to the spreading direction, showing that their formation is not a function of ridge obliquity. The absence of strike-slip earthquakes on the SWIR from 9° to 16° E confirms that these amagmatic segments are not an oblique form of transform. Although this supersegment formed recently, amagmatic accretionary segments can form long-term stable plate boundaries. Between 54° and 57° E on the SWIR, ref. 24 describes oblique non-transform discontinuities, 50 and 76 km long, connecting three orthogonal segments (Fig. 5). The orthogonal segments are typical magmatic ridge segments with large-amplitude negative MBA lows. The central magnetic anomaly amplitudes, however, decrease from their midpoints towards the oblique segments. These oblique segments are not discontinuities, however, as they extend the ridge for 31 and 44 km respectively, perpendicularly to the spreading direction. They have ESR values of 10.8 and 10.6 mm yr⁻¹, lower relief and weak MBA lows. These, then, are amagmatic accretionary ridge segments. Satellite gravity maps⁴⁶ show depressed topography between the elevated magmatic segments extending in the spreading direction across the Indian Ocean to the triple junction trace, indicating that these linked magmatic and amagmatic segments formed a stable plate boundary for >50 Myr (Fig. 5c).

High-angle normal faults and fissures equivalent to the intra-rift faults seen at highly oblique slow-spreading ridges are abundant along the SWIR oblique supersegment. However, these cut the rift boundary faults rather than interacting with them as at slow-spreading ridges. At the same time, the rift boundary faults are very unlike those bounding faster-spreading ridges. This demonstrates that, unlike oblique slow-spreading ridges, these two fault sets form and operate independently. The highly unfavourable orientations for brittle failure of the major boundary faults at the oblique amagmatic segments imply that strain is localized by pre-existing structures formed outside the brittle deformation field. The primary mode of accretion along the amagmatic segments appears to be emplacement of mantle peridotite directly to the rift valley floor along the axis of rifting and parallel to the ridge trend, with subsequent fault capture and block rotation to form the rift valley walls. Because these structures are parallel to the ridge trend, we suggest that failure is localized by the axis of intrusion of the asthenosphere into the lithosphere along the zone of lithospheric necking at the base of the plate. This plate-driven horst and graben tectonics is similar to recent finite-element models for rifting of thick lithosphere at ridges⁴⁷, and may provide a new analogue to explain features of early continental rifting.

At the magmatic accretionary segments the principal faults and the axes of volcanic eruption are orthogonal to the spreading direction. These features are consistent with brittle extensional plate failure, and hence from the top of the plate. Thus a fundamental difference between magmatic and amagmatic accretionary segments is that the latter effectively represent failure initiated from the base of the plate, whereas the former represent failure largely initiated from the top. Why transforms disappear and a new class of accretionary plate boundary structure replaces them is unknown, although it does seem to be due to a thicker, colder and therefore stronger lithosphere.

Whereas fast-spreading ridges lack deep rift valleys, and their segmentation patterns evolve rapidly, reflecting thin lithosphere, slow-spreading ridge segmentation patterns evolve slowly, and their tectonics reflect the interplay of brittle deformation processes at the top of the plate associated with continuous volcanism and the forces arising from extension of a thicker lithosphere. At ultraslow-spreading ridges, however, these latter forces dominate the tectonics, except where a local magmatic segment forms. Thus slow-spreading ridge tectonics is intermediate between that at fast- and ultraslow-spreading ridges. There are thus three distinct classes of ocean ridge: fast-, slow- and ultraslow-spreading, with two transitional intermediate classes, because ridge tectonics are not simply a

function of spreading rate, but also of ridge geometry, mantle composition and thermal structure. □

Received 23 December 2002; accepted 10 October 2003; doi:10.1038/nature02128.

- Solomon, S. in *Drilling the Oceanic Lower Crust and Mantle, JOI/USSAC Workshop Report* (ed. Dick, H. J. B.) 73–74 (Woods Hole Oceanographic Institution, Woods Hole, MA, 1989).
- Small, C. in *Faulting and Magmatism at Mid-Ocean Ridges* (eds Buck, W., Delaney, P. T., Karson, J. A. & Lagabriele, Y.) 1–26 (American Geophysical Union, Washington DC, 1998).
- Macdonald, K. C. in *The Western North Atlantic Region* (eds Vogt, P. & Tucholke, B.) 51–68 (Geological Society of America, Boulder, CO, 1986).
- Heezen, B. C. The rift in the ocean floor. *Sci. Am.* **203**, 99–106 (1960).
- Fornari, D. J., Haymon, R. M., Perfit, M. R. & Edwards, M. H. Geological characteristics and evolution of the axial zone on fast spreading mid-ocean ridges: formation of an axial summit trough along the East Pacific Rise, 9°–10°N 1998. *J. Geophys. Res.* **103**, 9827–9855 (1998).
- Small, C. & Sandwell, D. T. An abrupt change in ridge axis gravity with spreading rate. *J. Geophys. Res.* **94**, 17383–17317 (1989).
- Cochran, J. R. Systematic variation of axial morphology along the Southeast Indian Ridge. *Eos* **260**, 88 (1991).
- Phipps Morgan, J. & Chen, Y. Dependence of ridge-axis morphology on magma supply and spreading rate. *Nature* **364**, 706–708 (1993).
- Reid, I. & Jackson, H. R. Oceanic spreading rate and crustal thickness. *Mar. Geophys. Res.* **5**, 165–172 (1981).
- Jackson, H. R., Reid, I. & Falconer, R. K. H. Crustal structure near the Arctic mid-ocean ridge. *J. Geophys. Res.* **87**, 1773–1783 (1982).
- Bown, J. W. & White, R. S. Variation with spreading rate of oceanic crustal thickness and geochemistry. *Earth Planet. Sci. Lett.* **121**, 435–439 (1994).
- White, R. S., Minshull, T. A., Bickle, M. J. & Robinson, C. J. Melt generation at very slow-spreading oceanic ridges: constraints from geochemical and Geophysical data. *J. Petrol.* **42**, 1171–1196 (2001).
- Dick, H. J. B., Georgen, J. E., le Roex, A. P., Lin, J. & Madsen, J. A. The influence of ridge geometry on mantle melting at an ultra-slow spreading ridge. *Eos* **F919** (1998).
- Georgen, J. E., Lin, J. & Dick, H. J. B. Evidence from gravity anomalies for interactions of the Marion and Bouvet hotspots with the Southwest Indian Ridge: effect of transform offsets. *Earth Planet. Sci. Lett.* **187**, 283–300 (2002).
- Cochran, J. R., Kurras, G. J., Edwards, M. H. & Coakley, B. J. The Gakkel Ridge: Bathymetry, gravity anomalies and crustal accretion at extremely slow spreading rates. *J. Geophys. Res.* **2003**, doi:10.1029/2002JB001830 (2002).
- Macdonald, K. C., Scheirer, D. S., Carbotte, S. & Fox, P. J. It's only topography: Parts 1 & 2. *GSA Today* **3**, 1–35 (1993).
- Kuo, B.-Y. & Forsyth, D. W. Gravity anomalies of the ridge-transform system in the South Atlantic between 31 and 34.5°S: Upwelling centers and variations in crustal thickness. *Mar. Geophys. Res.* **10**, 205–232 (1988).
- Lin, J. & Phipps Morgan, J. The spreading rate dependence of three-dimensional mid-ocean ridge gravity structure. *Geophys. Res. Lett.* **19**, 13–16 (1992).
- Tolstoy, M., Harding, A. J. & Orcutt, J. A. Crustal thickness on the Mid-Atlantic Ridge: Bull's-eye gravity anomalies and focused accretion. *Science* **262**, 726–729 (1993).
- Magde, L. S. & Sparks, D. W. Three-dimensional mantle upwelling, melt generation, and melt migration beneath segment slow spreading ridges. *J. Geophys. Res.* **102**, 20571–20583 (1997).
- Dauteuil, O. & Brun, J. P. Oblique rifting in a slow spreading ridge. *Nature* **361**, 145–148 (1993).
- Clay, K. R. & White, M. J. Analogue modelling of orthogonal and oblique rifting. *Mar. Petrol. Geol.* **12**, 137–151 (1995).
- Jokat, W. *et al.* Geophysical evidence for reduced melt production on the Arctic ultraslow Gakkel mid-ocean ridge. *Nature* **423**, 962–965 (2003).
- Sauter, D. *et al.* The Southwest Indian Ridge between 49°15' E and 57°E: focused accretion and magma redistribution. *Earth Planet. Sci. Lett.* **192**, 303–317 (2001).
- Grindlay, N. R., Madsen, J. A., Rommevaux-Jestin, C. & Sclater, J. A different pattern of ridge segmentation and mantle Bouguer gravity anomalies along the ultra-slow spreading Southwest Indian Ridge (15°30' E to 25°E). *Earth Planet. Sci. Lett.* **161**, 243–253 (1998).
- le Roex, A. P., Dick, H. J. B. & Watkins, R. T. Petrogenesis of anomalous K-enriched MORB from the Southwest Indian Ridge: 11°53' E to 14°38' E. *Contrib. Mineral. Petrol.* **110**, 253–268 (1992).
- DeMets, C., Gordon, R. G., Argus, D. F. & Stein, S. Current plate motions. *Geophys. J. Int.* **101**, 425–478 (1990).
- DeMets, C., Gordon, R. G., Argus, D. F. & Stein, S. Effect of recent revisions to the geomagnetic reversal time scale on estimates of current plate motions. *Geophys. Res. Lett.* **21**, 2191–2194 (1994).
- Okino, K. *et al.* Preliminary analysis of the Knipovich Ridge segmentation: influence of focused magmatism and ridge obliquity on an ultraslow spreading system. *Earth Planet. Sci. Lett.* **202**, 279–288 (2002).
- Cochran, J. R. & Coakley, B. J. Morphology and segmentation of the Gakkel Ridge, Arctic Ocean, from SCICEX data. *Eos* **79**, F854 (1998).
- Brozena, J. M. *et al.* New aerogeophysical study of the Eurasia Basin and Lomonosov Ridge: Implications for basin development. *Geology* **31**, 825–828 (2003).
- Michael, P. J. *et al.* Magmatic and amagmatic seafloor generation at the ultraslow-spreading Gakkel Ridge, Arctic Ocean. *Nature* **423**, 956–961 (2003).
- Jackson, H. R. & Reid, I. Oceanic magnetic anomaly amplitudes: variation with sea-floor spreading rate and possible implications. *Earth Planet. Sci. Lett.* **63**, 368–378 (1983).
- Feden, R. H., Vogt, P. R. & Fleming, H. S. Magnetic and bathymetric evidence for the “Yermak Hot Spot” northwest of Svalbard in the Arctic Basin. *Earth Planet. Sci. Lett.* **44**, 18–38 (1979).
- Duckworth, G. L., Baggeroer, A. B. & Jackson, H. R. Crustal structure measurements near FRAMIL in the Pole Abyssal Plain. *Tectonophysics* **89**, 172–215 (1982).
- Michael, P. J. *et al.* The Arctic Mid-Ocean Ridge Expedition—AMORE 2001—Seafloor spreading at the top of the world. *Eos* **82**, F1097 (2001).
- Edwards, M. H. *et al.* Evidence of recent volcanic activity on the ultraslow-spreading Gakkel ridge. *Nature* **409**, 808–812 (2001).
- Coakley, B. J. & Cochran, J. R. Gravity evidence of very thin crust at the Gakkel Ridge (Arctic Ocean). *Earth Planet. Sci. Lett.* **162**, 81–95 (1998).

39. Vogt, P., Taylor, P. T., Kovacs, L. C. & Johnson, G. L. Detailed aeromagnetic investigation of the Arctic Basin. *J. Geophys. Res.* **84**, 1071–1089 (1989).
40. Standish, J., Dick, H., le Roex, A., Melson, W. & O'Hearn, T. Major and trace element geochemistry: Ultra-slow spreading SWIR (9°–25° E). *Eos* **83**, F1331 (2002).
41. Klein, E. M. & Langmuir, C. H. Global correlations of ocean ridge basalt chemistry with axial depth and crustal thickness. *J. Geophys. Res.* **92**, 8089–8115 (1987).
42. Salters, V. J. M. & Dick, H. J. B. Mineralogy of the mid-ocean-ridge basalt source from neodymium isotopic composition of abyssal peridotites. *Nature* **418**, 68–72 (2002).
43. Mahoney, J., Le Roex, A. P., Peng, Z., Fisher, R. L. & Natland, J. H. Southwestern limits of Indian Ocean ridge mantle and the origin of low $^{206}\text{Pb}/^{204}\text{Pb}$ mid-ocean ridge basalt: isotope systematics of the Central Southwest Indian Ridge (17°–50°E). *J. Geophys. Res.* **97**, 19771–19790 (1992).
44. Pertermann, M. & Hirschmann, M. M. Anhydrous partial melting experiments on MORB-like Éclogite: Phase relations, phase compositions and mineral-melt partitioning of major elements at 2–3 GPa. *J. Petrol.* **44**, 2173–2201 (2003).
45. Dick, H. J. B. & Natland, J. Late stage melt evolution and transport in the shallow mantle beneath the East Pacific Rise. In *Scientific Results* (eds Gillis, K., Mevel, C. & Allan, J.) 103–134 (Ocean Drilling Program, Texas A&M University, College Station, TX, 1996).
46. Sandwell, D. T. & Smith, W. H. F. Marine gravity anomaly from Geosat and ERS 1 satellite altimetry. *J. Geophys. Res.* **102**, 10039–10054 (1997).
47. Poliakov, A. N. B. & Buck, W. R. in *Faulting and Magmatism at Mid-Ocean Ridges* (eds Buck, W. R., Delaney, P. T., Karson, J. A. & Lagabriele, Y.) 305–324 (American Geophysical Union, Washington DC, 1998).
48. Phipps Morgan, J., Parmentier, E. M. & Lin, J. Mechanisms for the origin of mid-ocean ridge axial topography; implications for the thermal and mechanical structure of accreting plate boundaries. *J. Geophys. Res.* **92**, 12823–12836 (1987).
49. Sotin, C. & Parmentier, E. M. Dynamical consequences of compositional and thermal density stratification beneath spreading centers. *Geophys. Res. Lett.* **16**, 368–377 (1989).
50. Forsyth, D. W. in *Mantle Flow and Melt Generation at Mid-Ocean Ridges* (eds Phipps Morgan, J., Blackman, D. K. & Sinton, J. M.) 1–66 (American Geophysical Union, Washington DC, 1992).
51. Macnab, R. & Jakobsson, M. The International bathymetric chart of the Arctic Ocean (IBCAO): An improved morphological framework for oceanographic investigations. *Geophys. Res. Abstr.* **5**, 10909 (2003).

Acknowledgements The National Science Foundation funded this work. We thank the crew and scientific party of RV *Knorr* Voyage 162, Legs 7–9 and the AMORE scientific party for discussions, particularly P. Michael, G. Kurras, C. Langmuir, J. Snow and W. Jokat. We also wish to acknowledge J. Standish for coordinating rock description on both cruises. J. P. Morgan, R. Buck, G. Hirth and J. Standish provided reviews of the manuscript, and R. Searle, and J. Cann provided us with additional insights into ridge tectonics.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.J.B.D. (hdick@who.edu).

An atypical haem in the cytochrome b_6f complex

David Stroebel¹, Yves Choquet², Jean-Luc Popot¹ & Daniel Picot¹

¹Laboratoire de Physico-Chimie Moléculaire des Membranes Biologiques, CNRS/Université Paris 7, UMR 7099, ²Laboratoire de Physiologie Membranaire et Moléculaire du Chloroplaste, CNRS UPR 1261, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, F-75005 Paris, France

Photosystems I and II (PSI and PSII) are reaction centres that capture light energy in order to drive oxygenic photosynthesis; however, they can only do so by interacting with the multisubunit cytochrome b_6f complex. This complex receives electrons from PSII and passes them to PSI, pumping protons across the membrane and powering the Q-cycle. Unlike the mitochondrial and bacterial homologue cytochrome bc_1 , cytochrome b_6f can switch to a cyclic mode of electron transfer around PSI using an unknown pathway. Here we present the X-ray structure at 3.1 Å of cytochrome b_6f from the alga *Chlamydomonas reinhardtii*. The structure bears similarities to cytochrome bc_1 but also exhibits some unique features, such as binding chlorophyll, β -carotene and an unexpected haem sharing a quinone site. This haem is atypical as it is covalently bound by one thioether linkage and has no axial amino acid ligand. This haem may be the missing link in oxygenic photosynthesis.

Cytochrome b_6f (plastoquinone:plastocyanin oxidoreductase) and its partial homologue cytochrome bc_1 are thought to use a similar mechanism, the so-called Q-cycle^{1,2}, to couple electron transfer (from a hydroquinone to a cytochrome or plastocyanin) with proton translocation in order to generate a transmembrane electrochemical H^+ gradient ($\Delta\mu H^+$). Central to this mechanism is the two-electron oxidation of hydroplastoquinone coupled with the release of two protons at the Q_o site on one side of the membrane (the lumen in chloroplasts). One electron is transferred through a high-potential chain comprising an Fe_2S_2 cluster and a haem f , analogous to c_1 in cytochrome bc_1 , whereas the other is directed through haems b_L and b_H to reduce quinone at the Q_i site on the other side of the membrane (the stroma in chloroplasts); upon two successive reduction events at the Q_i site, two protons are taken up from the stroma. Thus $\Delta\mu H^+$ is generated through the release and uptake of protons on either side of the membrane.

Cytochromes bc_1 and b_6f have fundamental differences despite their mechanistic similarities. In oxygenic photosynthesis the electron pathway can switch from a linear mode, which starts with water and ends with NADPH, to a cyclic one, where electrons transferred from cytochrome b_6f to PSI, on the luminal face of the membrane, are fed back to b_6f on the stromal face, boosting ATP synthesis at the expense of reducing equivalents. This re-injection pathway is not understood. It may involve an NADH-plastoquinone reductase³, or may be mediated directly by ferredoxin, the ferredoxin:NADP⁺ oxidoreductase⁴, or by a biochemically unidentified G cytochrome^{5,6}. Furthermore, b_6f is thought to regulate state transitions by activating a protein kinase^{7,8}. A b_6f -specific proton-pumping mechanism has also been proposed⁹. Finally, two cofactors of unknown function, chlorophyll a and β -carotene, are part of cytochrome b_6f ¹⁰.

The X-ray structures of mitochondrial cytochrome bc_1 ^{11–14} have provided a structural framework for modelling cytochrome b_6f .

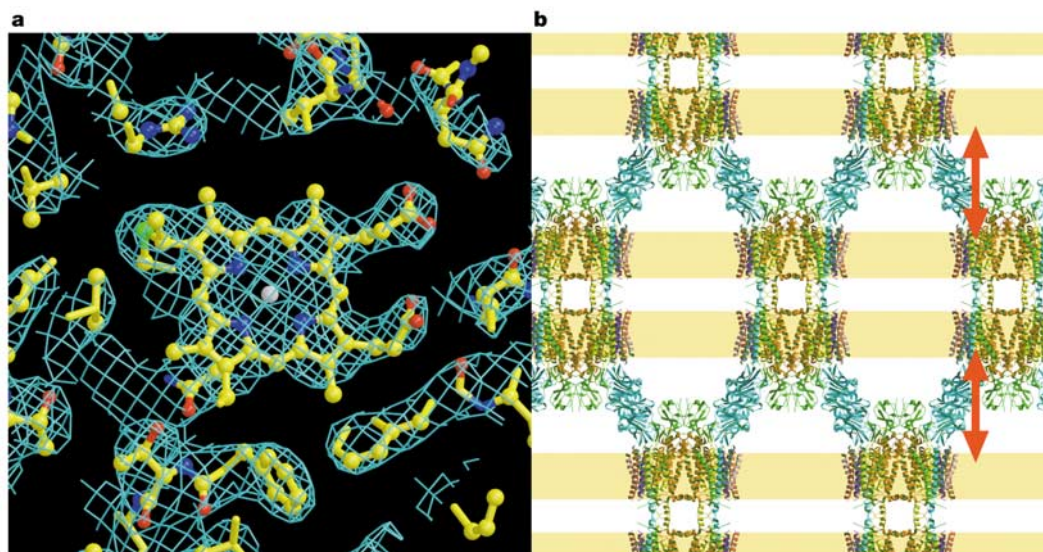


Figure 1 Crystallography of b_6f complex. **a**, Electron density of haem c_1 in the Q_1 site. ($2F_o - F_c$)-weighted map contoured at the 2σ level. Yellow, carbon; red, oxygen; blue, nitrogen; grey, iron. **b**, Crystal packing. A view normal to the molecular two-fold axis shows successive layers corresponding to stromal, transmembrane (in yellow) and

luminal (delimited with red arrows) regions of the complex. In the latter, contacts between layers involve exclusively cytochrome f . The thickness of the ‘luminal’ space (red arrow, 70 Å) is close to that measured in thylakoids⁵¹. The molecular two-fold axis is a crystallographic axis, and the crystal contains one monomer per asymmetric unit.

This has given useful insights into the Q_o site of b_6f , but puzzling results for Q_i . The X-ray structure (Table 1) of the b_6f complex described here explains these discrepancies and provides new clues for the understanding of oxygenic photosynthesis.

General polypeptide arrangement

The b_6f complex has a strong tendency to dissociate in detergent solution¹⁵. This problem has been overcome by the design of a rapid purification scheme using low detergent concentration, requiring less than a day from purification to crystallization assays. This procedure avoids adding exogenous lipids^{16,17}. The electron density map shown in Fig. 1 reveals all known subunits and cofactors plus several lipids.

The crystal structure shows a dimer (Figs 1b, 2 and 3) whose organization is similar to that of cytochrome bc_1 , as already suggested¹⁸. The extramembrane domains of cytochrome f and the Rieske protein lie on the luminal side. Although cytochromes f (ref. 19) and c_1 share no structural similarity, they occupy roughly equivalent positions in the two complexes. However, the c -type haem of cytochrome f is shifted by 16.3 Å relative to that of cytochrome c_1 . The extramembrane region of the Rieske protein is thought to be highly mobile^{12,13,20}. In the dimer the two cytochrome f subunits, which make no contact, form a bowl-like structure that, *in situ*, protects the movement of the two Rieske proteins from interactions in the narrow luminal space (Figs 1b and 2a). The electron density is well defined for the Fe_2S_2 -binding domain, but not for the other extramembrane domains, probably due to their mobility; however, with the exception of the link (residues 72–79) to the transmembrane helix, they can be localized²¹. The position of the Fe_2S_2 domain relative to the complex superimposes with that of the homologous domain in cytochrome bc_1 structures that binds stigmatellin¹⁴ (Fig. 3).

The transmembrane region

There are 13 transmembrane helices per monomer (Fig. 2b, c). Those of cytochrome b_6 and subunit IV exhibit a high degree of similarity to their homologues in the bc_1 cytochrome. The insertion of an additional residue (Val 190) between the two haem-liganding histidines on helix D produces limited local perturbations. The tilt of the helices relative to the membrane plane creates, as in cytochrome bc_1 , a groove at the dimer interface, thought to favour transfer of quinone between Q_o and Q_i . The single transmembrane helices of the Rieske protein and cytochrome f occupy equivalent locations to their counterparts in the bc_1 complex, but that of the Rieske protein is bent.

Table 1 Summary of data collection and refinement statistics	
Summary statistics	Values
Data collection	
Wavelength (Å)	1.0080
Unit cell dimensions (Å)	102.35, 171.15, 351.91
Maximum resolution (Å)	3.1
Unique reflections	56,687
Completeness (%)	99.6 (99.6)
Redundancy	4.3
R_{sym}	0.089 (0.39)
I/σ	6.8 (1.8)
Refinement	
Resolution range (last shell) (Å)	35–3.1 (3.21–3.10)
Number of atoms	7,820
$R_{\text{work}}/R_{\text{free}}$	0.22 (0.34)/0.26 (0.36)
Deviation from ideal geometry	
R.m.s. bonds (Å)/r.m.s. angle (°)	0.008/1.32
R.m.s. dihedral (°)/r.m.s. improper (°)	19.8/3.0
Average B factors (Å ²)	69.9

$R_{\text{sym}} = \sum |I - \langle I \rangle| / \sum I$, where I is the intensity. $R_{\text{work}} = \sum |F_o - F_c| / \sum F_o$, where F_o and F_c represent the observed and calculated structure factor amplitudes. R_{free} is calculated from about 5% of randomly chosen reflections, which are not used in the refinement. R.m.s., root mean square. Values in parentheses refer to the highest resolution shell.

The four small subunits PetG, PetL, PetM and PetN form an approximate four-helix bundle, which has no counterpart in cytochrome bc_1 . Conversely, helix H of cytochrome b and the small subunits su10, su7 and su11 (ref. 13) of bc_1 have no equivalent in cytochrome b_6f . Therefore, the two complexes feature different overall cross-sections (Fig. 2), with the small subunits of cytochrome b_6f occupying what in bc_1 is a large groove partially filled with specific lipids²².

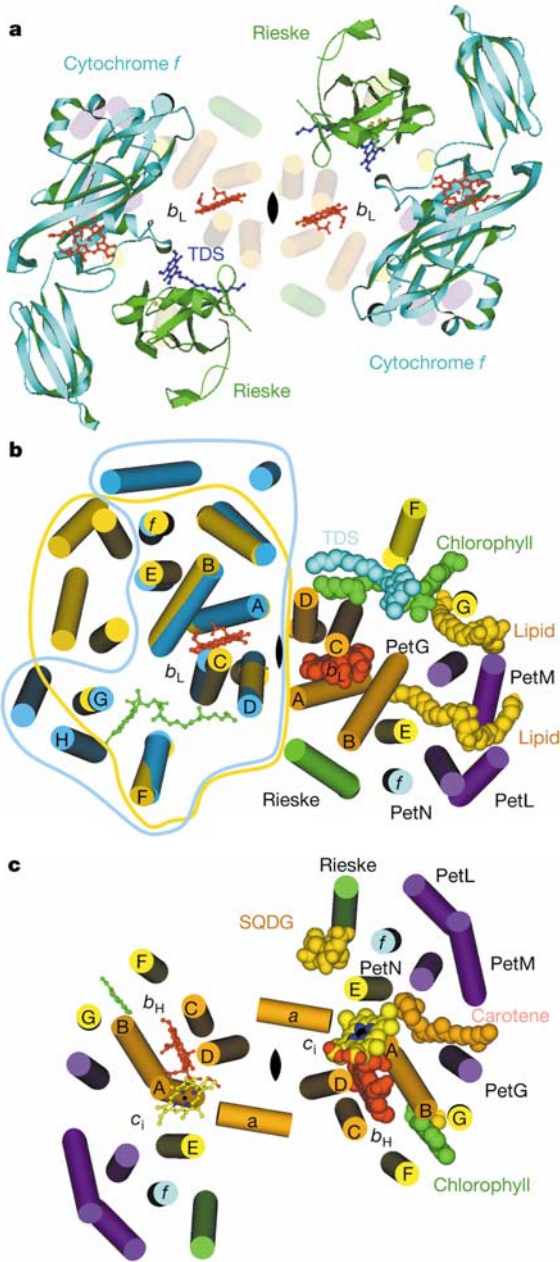


Figure 2 Three views of b_6f normal to the membrane. **a**, The lumen with cytochrome f and the Rieske protein. **b**, The transmembrane region viewed from the lumen. On the left, the transmembrane helices of cytochrome b_6f (yellow) are superimposed over those of cytochrome bc_1 (PDB entry 1BE3 (ref. 13); blue). The cross-section profile of the cytochrome bc_1 monomer is outlined in blue and that of cytochrome b_6f in yellow. On the right (cytochrome b_6 in orange, subunit IV in yellow and small subunits in purple), luminal cofactors and ligands are depicted in a space-filling representation. **c**, The stromal half of the transmembrane region viewed from the stroma (colours as in **b**). Helices are named according to cytochrome bc_1 terminology¹²: A, B, C and D for cytochrome b_6 ; E, F and G for subunit IV; helix H is present only in cytochrome b .

Chlorophyll

Chlorophyll *a* is a component of the b_6f complex¹⁰. Its chlorin ring is located between helices F and G of subunit IV (Fig. 2b, c), at the centre of the transmembrane region, in a space left accessible by the absence of a homologous helix H, which is present in mitochondrial cytochrome *b*. On both faces of the ring there are weak electron densities that can be attributed to lipids or other lipophilic compounds. No clear axial ligand can be seen, although electron density extends away from the chlorin plane, towards the carbonyl group of Gly 136 of subunit IV. The distance is too large (4.7 Å) for direct coordination, suggesting an intervening water molecule. The phytyl chain ends within the Q_o pocket.

β -carotene

β -carotene co-purifies with cytochrome b_6f ^{10,23}. One cyclohexene ring is deeply embedded not far (approximately 6.2 Å) from haem c_1 (see below), whereas the other ring disappears from the electron density map at the protein–lipid interface (between PetG and PetM). A good fit to the electron density is observed with a 9-*cis*- β -carotene, in agreement with Raman spectra²⁴. The β -carotene molecule seems too distant (14 Å) from the chlorophyll to efficiently quench singlet oxygen, in agreement with spectroscopic data²⁵, but it could interact with the Q_i site. The presence of additional carotenoids interacting with the chlorophyll *in situ* cannot be ruled out, as the edge of the chlorophyll is exposed to the lipid phase.

Specific features of cytochrome b_6f —the plastocyanin-binding site of cytochrome *f* (ref. 26), the small subunits helix bundle, the chlorophyll and the β -carotene—are all facing in the same direction out of the complex. This suggests that they could interact with other components of the photosynthetic apparatus, such as PSI or a kinase. The chlorophyll might serve as a sensor that connects the interacting partner with the Q_o site. The β -carotene might have a similar role at the Q_i site; however, other hypotheses¹⁰ cannot be ruled out.

Endogenous lipids

On the stromal side lies a sulpholipid (sulphoquinovosyldiacylglycerol; SQDG) with partially visible fatty acyl chains (Fig. 2c). The

head group electron density is well defined, and the sulphur atom produces a peak in an anomalous Fourier map. The sulphonate group interacts with Lys 272 at the stromal end of the transmembrane helix of cytochrome *f*. This lysine is involved in regulation of cytochrome *f* biosynthesis²⁷. The luminal side features two lipids (probably monogalactosyldiacylglycerol; Fig. 2b)—two acyl chains with well-defined electron densities are embedded in the protein, in contact with β -carotene in one case and chlorophyll *a* in the other. The other acyl chains, partially disordered, lie at the lipid–protein interface.

Q_o and tridecylstigmatellin-binding sites

The luminal face harbours the high-potential chain and haem b_L of the low-potential chain. The Q_o site (Fig. 4), where hydroplastoquinone oxidation and proton release take place, is reminiscent of the bc_1 Q_o site; that is, it exhibits the same bifurcated volume, with its lobe distal from haem b_L occupied by the head group of the inhibitor tridecylstigmatellin (TDS). The entrance to the active site is partially occupied by the phytyl chain of the chlorophyll. Less bulky side chains at the opposite wall of the cleft compensate for the presence of the phytyl chain. This reorganization imposes onto the tail of TDS a shift of about 4 Å relative to that of stigmatellin in cytochrome bc_1 ²². The His 155 residue of the Rieske protein is within hydrogen-bonding distance of the carbonyl of the chromone ring, the suggested path for the release of one proton¹². Differences exist in the lobe proximal to haem b_L . For instance, the side chain of Glu 78 of subunit IV (Glu 272 in yeast cytochrome *b*) has a poorly resolved electron density, suggesting mobility, whereas in bc_1 this residue has been proposed to have an important role in the release of the other proton²⁸; its substitution by polar or basic residues in *C. reinhardtii* suggests that it is not essential for b_6f activity²⁹. In the

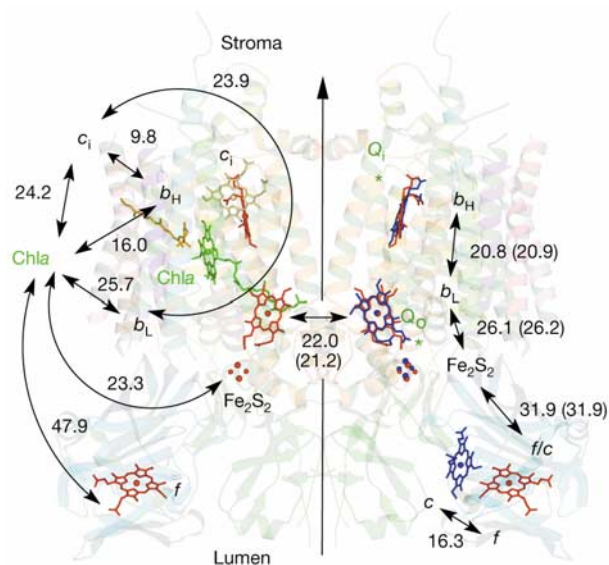


Figure 3 Cofactors in the dimer. Cytochrome b_6f 's haems (red, haem c_1 in yellow), Fe_2S_2 (red) and chlorophyll (green) are represented as viewed along the membrane plane. On the right, the cofactors of cytochrome bc_1 (blue) are superimposed (PDB entry 1KB9; ref. 22). Distances between metal centres (in Å) are indicated for cytochrome b_6f and are in parentheses for cytochrome bc_1 . The angle between the planes of haem c_1 and b_H is 74°; the iron of haem c_1 is 3.9 Å closer to the stroma than that of b_H .

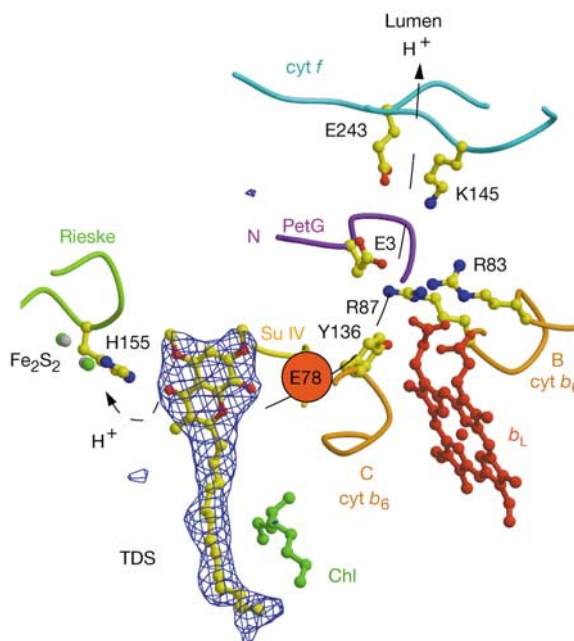


Figure 4 Q_o site with inhibitor. TDS is shown with a $(F_o - F_c)$ -annealed omit map contoured at 5 σ . Residue His 155 of the Rieske protein is within hydrogen-bonding distance of the carbonyl of TDS. To the right of TDS is the region around the lobe proximal to haem b_L , with the beginning of the low-potential chain and a possible channel for the second released proton. Charged residues are indicated (they are not hydrogen-bonded; the distance between two neighbouring groups is about 5–6 Å, suggesting intervening water molecules). The conformation of cytochrome b_6 Tyr 136 differs from that in cytochrome bc_1 . Su IV, subunit IV.

proximal lobe the main chain undergoes, between residues 58 and 63 of subunit IV, an 8 Å shift relative to the mitochondrial complex, making space for the amino terminus of PetG and modifying the distribution of charged residues, implying different paths for proton escape³⁰ (Fig. 4).

An additional haem on the stromal face at the Q_i site

The Q_i site is located on the stromal face, where proton uptake and reduction of plastoquinone (PQ) take place. It opens onto the hydrophobic groove at the dimer interface, as in cytochrome *bc*₁, but the active site is different. Indeed, a large, flat electron density, with two lateral extensions and an electron dense region at its centre (Fig. 1a), blocks access to haem *b*_H. This density can be attributed to a haem on the following grounds: (1) the strong density at its centre originates from an iron atom, as shown by anomalous Fourier maps calculated from data collected at several wavelengths (1.009, 1.215 and 1.738 Å), and by a difference Fourier map between two data sets taken at and near the absorption edge of the iron K α transition (Fig. 5a). (2) The general shape of the electron density resembles

that of the three known haems, with well-defined densities for the propionates (Fig. 1). (3) One vinyl group of the proposed haem is within bonding distance of the sulphur atom of Cys 35 from cytochrome *b*₆, consistent with the presence of a thioether bridge. The additional haem, therefore, must belong to the *c* group, and will henceforth be called haem *c*_i; however, the consensus sequence CXnCH or AXnCH is lacking. Although uncommon, haem attachment by a single thioether bridge has been reported previously³¹. (4) Pyridine haemochromogen spectra (data not shown) are consistent with the presence of an additional haem with one thioether bridge^{31,32}.

The discovery of an additional haem in an extensively studied complex was not anticipated. Nevertheless, there were clues to the presence of a covalently bound haem: (1) it explains the peroxidase activity of cytochrome *b*₆ in SDS–polyacrylamide gel electrophoresis gels³³, which is not observed with cytochrome *bc*₁; (2) it accounts for the excess mass of cytochrome *b*₆ in mass spectrometry experiments³⁴; and (3) it is consistent with the involvement of four nuclear genes in haem binding to *C. reinhardtii*'s cytochrome *b*₆ (ref. 35).

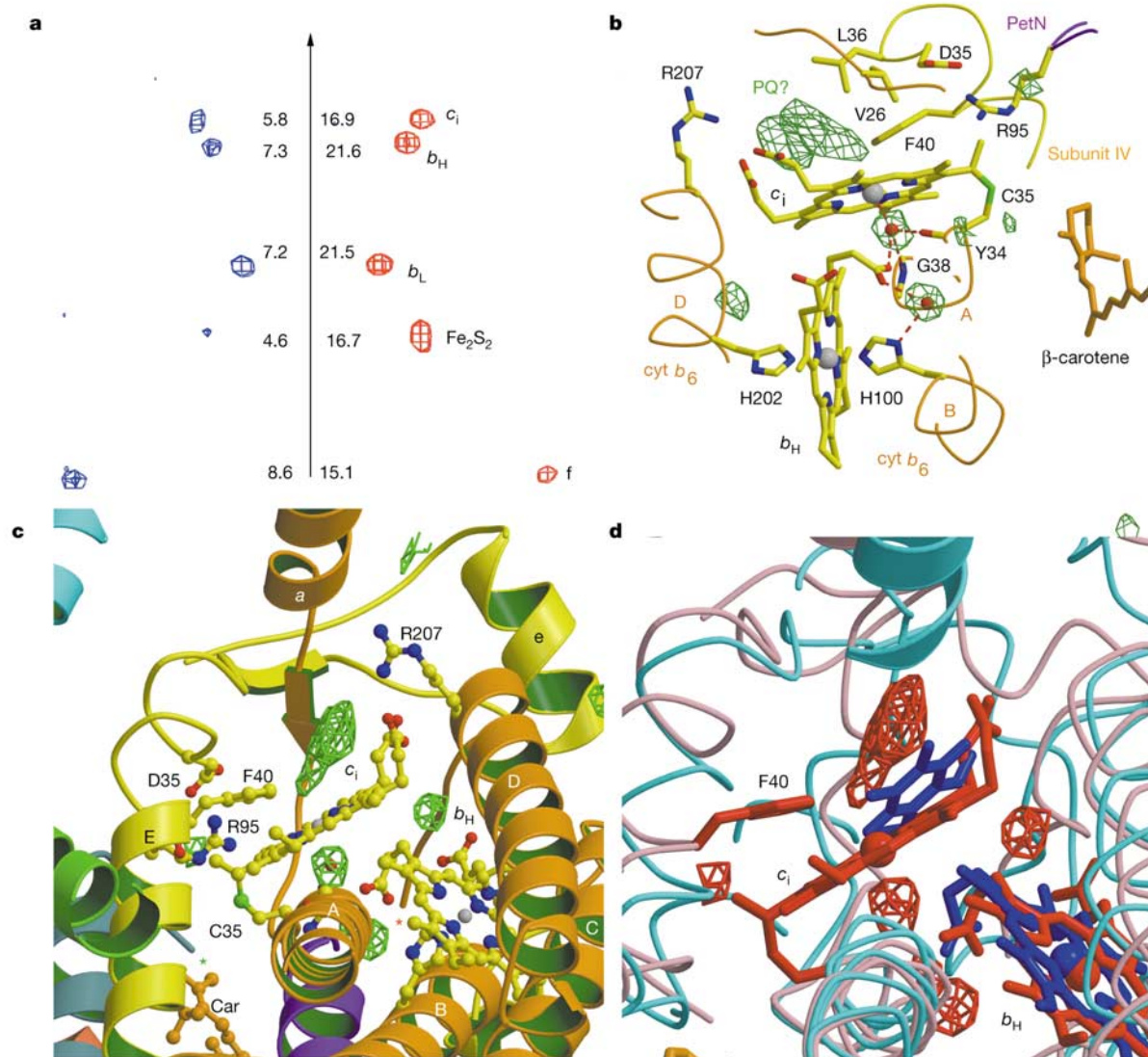


Figure 5 Q_i site. **a**, Irons in the dimer. Left, anomalous Fourier difference map ($\Delta F'$) at the iron edge (1.737–1.7408 Å) contoured at 4.0 σ ; right, anomalous map ($\Delta F''$) at 1.7389 Å contoured at 8.5 σ . The arrow represents the two-fold axis. Peak height units are in σ . **b**, Haem *c*_i environment. The ($F_o - F_c$) map contoured at 5 σ (in green) shows potential ligand and water molecules around haem *c*_i (also visible in **c** (green) and **d** (red)). **c**, Q_i site

from the dimer interface. Red and green asterisks mark the C termini of cytochrome *b*₆ and PetN, respectively. **a**, N-terminal transverse helix of cytochrome *b*₆; Car, β -carotene. **d**, Superposition of cytochrome *b*₆*f* (red) and cytochrome *bc*₁ (PDB entry 1KB9; ref. 22) (blue) in the Q_i region. In cytochrome *bc*₁, haem *b*_H and an ubiquinone are depicted.

Axial ligands

The atypical character of haem c_i is accentuated by the unusual coordination of the iron atom, which does not directly interact with any amino acid. On its face close to b_H , Fourier differences maps ($F_o - F_c$) reveal the presence of an extra density, compatible with either a water molecule or an hydroxide ion (Fig. 5b), located within hydrogen-bonding distance of the carboxyl group of one propionate of haem b_H , the main-chain carbonyl of Tyr 34 and the amide of Gly 38 of cytochrome b . The same propionic carboxyl in turn is within hydrogen-bonding distance of another buried water molecule—also present in cytochrome bc_1 (ref. 22)—which itself can form a hydrogen bond with the N δ atom of His 100, an axial ligand of haem b_H . The other face of haem c_i lines the active site pocket, where a weak electron density could be accounted for by either PQ or another ligand (Fig. 5b, c).

A haem with no axial amino acid ligand has never been observed in wild-type proteins, but it has been created by mutagenesis of myoglobin³⁶ or peroxidases³⁷. Haem c_i should belong to the c' group (that is high-spin), whose broad spectral bands in the visible region of the spectrum ought to be dominated by those of the low-spin haems and other chromophores of the cytochrome b_6f . The agreement between linear dichroism spectra³⁸ and the orientation of haems f , b_L and b_H in the present structure bears witness to the proper assignment of their absorption bands. Electron paramagnetic resonance (EPR) spectra of purified b_6f samples do feature high-spin components³⁹, which would be consistent with the presence of a c' -type haem. However, as haems b_H and c_i are in van der Waals contact (Fig. 3), both haems could be EPR-silent.

Q_i site

The polypeptide chain around Q_i has the same topology as in cytochrome bc_1 , but the active site has to accommodate a substrate and haem c_i , thus forming a different active site and explaining the lack of inhibitory effect of antimycin on cytochrome b_6f . Indeed, the amphipathic helix a of cytochrome b_6 , which lies at the dimer interface parallel with the membrane plane (Fig. 2c), is shifted by about 6 Å towards the stroma relative to that in cytochrome bc_1 (Fig. 5c, d). This creates, adjacent to haem c_i , a pocket that is large enough for a substrate to interact with the haem plane, and, possibly, with the iron itself. This configuration differs from all known quinone-binding sites. Residue Phe 40 from subunit IV is in van der Waals' contact with the haem plane. Its role may be to control interaction of ligands with the haem iron. Near the entrance of the active site cavity one propionate of haem c_i is within hydrogen-bonding distance of Arg 207 at the carboxy-terminal end of helix D (His 202 in cytochrome bc_1). This residue is in a favourable position to interact with a quinone in the pocket.

Haem c_i and phylogenesis of cytochrome b_6 -type complexes

The presence of an additional haem in the Q_i site ought to be a general feature of all cytochrome b_6f -like complexes. Indeed, Cys 35 of cytochrome b_6 —to which haem c_i is covalently bound—as well as other amino acids of the Q_i pocket are well conserved in all chloroplast and cyanobacterial cytochrome b_6f complexes. No cysteine is present at that position in mitochondrial and most bacterial bc_1 cytochromes. It is, however, encountered in menaquinol:cytochrome c oxidoreductase from some members of the Firmicutes phylum (but not the Chlorobiaceae), such as *Bacillus subtilis*⁴⁰, *Heliobacillus mobilis* or *Heliobacterium gestii*; the latter two are photosynthetic and harbour a PSI-like reaction centre. The same pattern is found in this complex from these species: cytochrome b is split into cytochrome b_6 and subunit IV, subunit IV is predicted to feature three transmembrane helices, and the two haem-liganding histidines of helix D are separated by 14 amino acids⁴¹. It is not clear from the structure whether there is a necessity for having a split cytochrome b , but it could be a constraint from the

haem c_i insertion machinery. Firmicutes and cyanobacteria are not that distant in evolutionary terms. Photosynthetic organisms that potentially harbour haem c_i all have a PSI-like reaction centre, and could function in a cyclic mode of electron transfer. Therefore, it is tempting to assign c_i to that role, even though direct evidence is lacking. In the non-photosynthetic Firmicutes there is a seven-residue insertion between residues 31 and 32 of cytochrome b_6 , which might modify or obstruct a potential path for electron re-injection from the stroma site to Q_i (see below). There is, however, one noticeable outlier; the evolutionarily distant proteobacterium *Xylella fastidiosa* exhibits a cysteine equivalent to Cys 35 and an unsplit cytochrome b , but it is unclear whether a haem c_i is present, as the amino acids that would surround it have a mixed b_6/b pattern.

Functional implications

Haem c_i blocks quinone access to b_H . This intermediate position suggests its participation in electron flow from b_H to PQ, either as a 'wire' between b_H and PQ, or as a redox carrier. Indeed, in contrast with the bc_1 complex, light-induced electron injection in the low-potential chain of the b_6f complex under oxidizing conditions results in the reduction of only a fraction (up to 0.7) of b_H , without the corresponding formation of a detectable semiquinone species⁴² (note however that the proximity of haem c_i would influence the EPR signal of a hypothetical semiquinone). Electron sharing between b_H and c_i might explain the fate of the missing electrons. In keeping with this hypothesis, ref. 5 established the existence of a carrier called G in *Chlorella sorokiniana* whose potential would be close to that of b_H and whose characteristics agree with the present structure. At present it needs to be demonstrated whether haem c_i is G. The accessibility of haem c_i from the stroma makes it a good candidate to participate in cyclic electron flow around PSI. Direct access to the Q_i site from the stroma is only prevented by the presence of Asp 35 (from subunit IV) and Arg 95 (from PetN) (Fig. 5b). Therefore, the existence of haem c_i brings further support to the idea that the re-injection of electrons occurs through cytochrome b_6f , even if details of the pathway from PSI remain unclear.

Untangling the interactions between the two Q_i haems will require close examination. Because a propionate of b_H interacts with the haem iron of haem c_i through an intervening water molecule or hydroxide ion, and because the pK of propionate is influenced by the redox state of b_H , the latter can modulate the redox state of haem c_i and, possibly, change the protonation state of the intervening molecule. This may be the heart of a proton pump.

The structure of a cyanobacterial cytochrome b_6f has been published during the revision of the present article⁴³. Among others, differences exist concerning the mode of TDS binding, and hence the position of the Rieske protein, and the assignment of subunits PetG, PetL and PetN. □

Methods

Protein purification

Chlamydomonas reinhardtii was genetically modified in order to add a six-histidine tag at the C terminus of cytochrome f (ref. 27). Thylakoid membranes were solubilized with dodecyl- β -maltoside (DDM, Anatrace). After centrifugation, the supernatant was loaded onto an anion-exchange column (Source 30Q, Amersham), equilibrated with 20 mM Tris-HCl pH 8 and 1 mM DDM. Elution was carried out with 220 mM NaCl in the presence of 0.5 mM DDM. The eluate was loaded onto a 1-ml nickel column (HiTrap chelating, Amersham) equilibrated with 20 mM Tris-HCl pH 8, 250 mM NaCl, 0.3 mM DDM. The detergent concentration was reduced to 0.2 mM and the protein eluted with 300 mM imidazole. After desalting, the protein was concentrated to 10 g l⁻¹ by ultrafiltration (Microcon YM100, Millipore). Pyridine haemochromogen spectra were measured according to ref. 32. TDS was synthesized according to ref. 44.

Crystallization

Crystals were obtained by vapour diffusion (hanging drops) after addition of TDS (2:1 molar ratio of inhibitor to protein), the reservoir composition being 40 mM Tris-HCl pH 8, 40 mM NaCl, 0.2 mM DDM, 30% (w/v) glycerol and 25% (w/v) polyethylene glycol monomethylether (molecular mass 350 Da) (Fluka).

Crystallography

Data were collected at 70–100 K on the synchrotron beamline BM30A (CRG FIP) of the ESRF, and processed with the CCP4 package⁴⁵. The space group of the crystal is I222. The solvent content is 82% of the crystal volume. Phases were obtained by multiple isomorphous replacement (see Supplementary Information) using crystals soaked in either 0.2 mM HgCl₂ or 2 mM TbNO₃. Heavy atoms were localized using SHELXD. Phases were refined with MLPHARE from the CCP4 package and improved by solvent modification using DM. The resulting electron density map allowed localizing the transmembrane helices and cytochrome *f*, whose extramembrane domain was replaced with the model of the soluble fragment from *C. reinhardtii* (Protein Data Bank (PDB) file 1E2W)⁴⁶. Heavy-atom phases were combined with model phases of cytochrome *f* with SIGMAA and solvent flattened with DM. The model was built with the program O (ref. 47) (Hg binding to Cys 7 of PetG and Cys 81 of PetN confirmed the localization of these small subunits) and the structure was refined with CNS⁴⁸. Figures were drawn with MOLSCRIPT⁴⁹ and RASTER3D⁵⁰.

Received 22 August; accepted 28 October 2003; doi:10.1038/nature02155.

- Mitchell, P. Possible molecular mechanisms of the protomotive function of cytochrome systems. *J. Theor. Biol.* **62**, 327–367 (1976).
- Crofts, A. & Wraight, C. A. The electrochemical domain of photosynthesis. *Biochim. Biophys. Acta* **726**, 149–185 (1983).
- Sazanov, L. A., Burrows, P. A. & Nixon, P. J. The plastid *ndh* genes code for an NADH-specific dehydrogenase: isolation of a complex I analogue from pea thylakoid membranes. *Proc. Natl Acad. Sci. USA* **95**, 1319–1324 (1998).
- Zhang, H., Whitelegge, J. P. & Cramer, W. A. Ferredoxin:NADP⁺ oxidoreductase is a subunit of the chloroplast cytochrome *b₆f* complex. *J. Biol. Chem.* **276**, 38159–38165 (2001).
- Lavergne, J. Membrane potential-dependent reduction of cytochrome *b₆* in an algal mutant lacking photosystem I centers. *Biochim. Biophys. Acta* **725**, 25–33 (1983).
- Joliot, P. & Joliot, A. The low potential electron transfer chain in cytochrome *b₆f* complex. *Biochim. Biophys. Acta* **933**, 319–333 (1988).
- Depege, N., Bellafiore, S. & Roach, J.-D. Role of chloroplast protein kinase Stt7 in LHCII phosphorylation and state transition in *Chlamydomonas*. *Science* **299**, 1572–1575 (2003).
- Wollman, F.-A. State transitions reveal the dynamics and flexibility of the photosynthetic apparatus. *EMBO J.* **20**, 3623–3630 (2001).
- Joliot, P. & Joliot, A. *In vivo* analysis of the effect of dicyclohexylcarbodiimide on electron and proton transfers in cytochrome *b₆f* complex of *Chlorella sorokiniana*. *Biochemistry* **37**, 10404–10410 (1998).
- Pierre, Y. *et al.* On the presence and role of a molecule of chlorophyll *a* in the cytochrome *b₆f* complex. *J. Biol. Chem.* **272**, 21901–21908 (1997).
- Xia, D. *et al.* Crystal structure of the cytochrome *bc₁* complex from bovine heart mitochondria. *Science* **277**, 60–66 (1997).
- Zhang, Z. *et al.* Electron transfer by domain movement in cytochrome *bc₁*. *Nature* **392**, 677–684 (1998).
- Iwata, S. *et al.* Complete structure of the 11-subunit bovine mitochondrial cytochrome *bc₁* complex. *Science* **281**, 64–71 (1998).
- Hunte, C., Koepke, J., Lange, C., Rossmanith, T. & Michel, H. Structure at 2.3 Å resolution of the cytochrome *bc₁* complex from the yeast *Saccharomyces cerevisiae* co-crystallized with an antibody Fv fragment. *Struct. Fold Des.* **8**, 669–684 (2000).
- Breyton, C., Tribet, C., Olive, J., Dubacq, J.-P. & Popot, J.-L. Dimer to monomer conversion of the cytochrome *b₆f* complex. Causes and consequences. *J. Biol. Chem.* **272**, 21892–21900 (1997).
- Pierre, Y., Breyton, C., Kramer, D. & Popot, J.-L. Purification and characterization of the cytochrome *b₆f* complex from *Chlamydomonas reinhardtii*. *J. Biol. Chem.* **270**, 29342–29349 (1995).
- Zhang, H., Kurisu, G., Smith, J. L. & Cramer, W. A. A defined protein-detergent-lipid complex for crystallization of integral membrane proteins: The cytochrome *b₆f* complex of oxygenic photosynthesis. *Proc. Natl Acad. Sci. USA* **100**, 5160–5163 (2003).
- Breyton, C. The cytochrome *b₆f* complex: structural studies and comparison with the *bc₁* complex. *Biochim. Biophys. Acta* **1459**, 467–474 (2000).
- Martinez, S. E., Huang, D., Szczepaniak, A., Cramer, W. A. & Smith, J. L. Crystal structure of chloroplast cytochrome *f* reveals a novel cytochrome fold and unexpected heme ligation. *Structure* **2**, 95–105 (1994).
- Darrouzet, E., Moser, C. C., Dutton, P. L. & Daldal, F. Large scale domain movement in cytochrome *bc₁*: a new device for electron transfer in proteins. *Trends Biochem. Sci.* **26**, 445–451 (2001).
- Carrell, C. J., Zhang, H., Cramer, W. A. & Smith, J. L. Biological identity and diversity in photosynthesis and respiration: structure of the lumen-side domain of the chloroplast Rieske protein. *Structure* **5**, 1613–1625 (1997).
- Lange, C., Nett, J. H., Trumpower, B. L. & Hunte, C. Specific roles of protein-phospholipid interactions in the yeast cytochrome *bc₁* complex structure. *EMBO J.* **20**, 6591–6600 (2001).
- Zhang, H., Huang, D. & Cramer, W. A. Stoichiometrically bound β-carotene in the cytochrome *b₆f* complex of oxygenic photosynthesis protects against oxygen damage. *J. Biol. Chem.* **274**, 1581–1587 (1999).
- Yan, J., Liu, Y., Mao, D., Li, L. & Kuang, T. The presence of 9-cis-β-carotene in cytochrome *b₆f* complex from spinach. *Biochim. Biophys. Acta* **1506**, 182–188 (2001).
- Peterman, E. J. *et al.* Fluorescence and absorption spectroscopy of the weakly fluorescent chlorophyll *a* in cytochrome *b₆f* of *Synechocystis* PCC6803. *Biophys. J.* **75**, 389–398 (1998).
- Hart, S. E., Schlarb-Ridley, B. G., Delon, C., Bendall, D. S. & Howe, C. J. Role of charges on cytochrome *f* from the cyanobacterium *Prothium laminosum* in its interaction with plastocyanin. *Biochemistry* **42**, 4829–4836 (2003).

- Choquet, Y., Zito, F., Wostrikoff, K. & Wollman, F.-A. Cytochrome *f* translation in *Chlamydomonas* chloroplast is autoregulated by its carboxyl-terminal domain. *Plant Cell* **15**, 1443–1454 (2003).
- Crofts, A. R. *et al.* Pathways for proton release during ubiquinol oxidation by the *bc₁* complex. *Proc. Natl Acad. Sci. USA* **96**, 10021–10026 (1999).
- Zito, F., Finazzi, G., Joliot, P. & Wollman, F.-A. Glu78, from the conserved PEWY sequence of subunit IV, has a key function in cytochrome *b₆f* turnover. *Biochemistry* **37**, 10395–10403 (1998).
- Finazzi, G. Redox-coupled proton pumping activity in cytochrome *b₆f*, as evidenced by the pH dependence of electron transfer in whole cells of *Chlamydomonas reinhardtii*. *Biochemistry* **41**, 7475–7482 (2002).
- Mukai, K. *et al.* An atypical heme-binding structure of cytochrome *c₁* of *Euglena gracilis* mitochondrial complex III. *Eur. J. Biochem.* **178**, 649–656 (1989).
- Berry, E. A. & Trumpower, B. L. Simultaneous determination of hemes *a*, *b*, and *c* from pyridine hemochrome spectra. *Anal. Biochem.* **161**, 1–15 (1987).
- Hurt, E. & Hauska, G. A cytochrome *flb₆* complex of five polypeptides with plastoquinol-plastocyanin-oxidoreductase activity from spinach chloroplasts. *Eur. J. Biochem.* **117**, 591–595 (1981).
- Whitelegge, J. P., Zhang, H., Aguilera, R., Taylor, R. M. & Cramer, W. A. Full subunit coverage liquid chromatography electrospray ionization mass spectrometry (LCMS⁺) of an oligomeric membrane protein: cytochrome *b₆f* complex from spinach and the cyanobacterium *Mastigocladus laminosus*. *Mol. Cell Proteom.* **1**, 816–827 (2002).
- Kuras, R. *et al.* Molecular genetic identification of a pathway for heme binding to cytochrome *b₆*. *J. Biol. Chem.* **272**, 32427–32435 (1997).
- Pond, A. E. *et al.* Assignment of the heme axial ligand(s) for the ferric myoglobin (H93G) and heme oxygenase (H25A) cavity mutants as oxygen donors using magnetic circular dichroism. *Biochemistry* **38**, 7601–7608 (1999).
- Hirst, J. *et al.* Replacement of the axial histidine ligand with imidazole in cytochrome *c* peroxidase. 2. Effects on heme coordination and function. *Biochemistry* **40**, 1274–1283 (2001).
- Schoepf, B., Chabaud, E., Breyton, C., Vermiglio, A. & Popot, J.-L. On the spatial organization of hemes and chlorophyll in cytochrome *b₆f*. A linear and circular dichroism study. *J. Biol. Chem.* **275**, 5275–5283 (2000).
- Nitschke, W. & Hauska, G. On the nature of the *g* = 6 EPR signal in isolated cytochrome *b₆f* complex from spinach chloroplasts. *Biochim. Biophys. Acta* **892**, 314–319 (1987).
- Yu, J. & Le Brun, N. E. Studies of the cytochrome subunits of menaquinone:cytochrome *c* reductase (*bc* complex) of *Bacillus subtilis*. Evidence for the covalent attachment of heme to the cytochrome *b* subunit. *J. Biol. Chem.* **273**, 8860–8866 (1998).
- Schütz, M. *et al.* Early evolution of cytochrome *bc* complexes. *J. Mol. Biol.* **300**, 663–675 (2000).
- Hope, A. B. The chloroplast cytochrome *b₆f* complex: a critical focus on function. *Biochim. Biophys. Acta* **1143**, 1–22 (1993).
- Kurisu, G., Zhang, H., Smith, J. L. & Cramer, W. A. Structure of the cytochrome *b₆f* complex of oxygenic photosynthesis: Tuning the cavity. *Science*, 2 October 2003 (doi:10.1126/science.1090165).
- Höfle, G., Kunze, B., Zorzin, C. & Reichenbach, H. Antibiotica aus gleitende Bakterien XIII: Stigmatellin A und B—zwei neue Antibiotika aus *Stigmatella aurantiaca* (Myxobacterales). *Liebigs Ann. Chem.* **12**, 1883–1904 (1984).
- The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Sainz, G. *et al.* Interruption of the internal water chain of cytochrome *f* impairs photosynthetic function. *Biochemistry* **39**, 9164–9173 (2000).
- Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Kraulis, P. J. A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
- Merritt, E. & Bacon, D. Raster3D photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524 (1997).
- Murakami, S. & Packer, L. Protonation and chloroplast membrane structure. *J. Cell Biol.* **47**, 332–351 (1970).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the local contacts of beamline BM30 (ESRF, CRG FIP) J.-L. Ferrer, R. Kahn and P. Carpentier for assistance in data collection; C. Lebreton for membrane preparations; P. Hervé and P. Fellmann for the synthesis of TDS; Y. Pierre, F. Zito and C. Breyton for biochemical expertise; P. Joliot, J. Lavergne, F. Daldal, F. Baymann, W. Nitschke and G. Ajlani for discussions; and C. Breyton, G. Finazzi and A. A. Pascal for discussions and critical reading of the manuscript. This work was supported by the CNRS and Paris-7 University.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.P. (Daniel.Picot@ibpc.fr) or J.-L.P. (Jean-Luc.Popot@ibpc.fr). Coordinates and structure factors have been deposited with the PDB under accession code 1Q90.

The formation of the Kuiper belt by the outward transport of bodies during Neptune's migration

Harold F. Levison^{1,2} & Alessandro Morbidelli²

¹Southwest Research Institute, Boulder, Colorado 80302, USA

²Observatoire de la Côte d'Azur, 06304 Nice Cedex 4, France

The 'dynamically cold Kuiper belt' consists of objects on low-inclination orbits between ~ 40 and ~ 50 AU from the Sun. It currently contains material totalling less than a tenth the mass of the Earth^{1,2}, which is surprisingly low because, according to accretion models^{3,4}, the objects would not have grown to their present size unless the cold Kuiper belt originally contained tens of Earth masses of solids. Although several mechanisms have been proposed to produce the observed mass depletion, they all have significant limitations⁵. Here we show that the objects currently observed in the dynamically cold Kuiper belt were most probably formed within ~ 35 AU and were subsequently pushed outward by Neptune's 1:2 mean motion resonance during its final phase of migration. Combining our mechanism with previous work^{6,7}, we conclude that the entire Kuiper belt formed closer to the Sun and was transported outward during the final stages of planet formation.

Since the discovery of the first Kuiper-belt object ten years ago⁸, a picture has emerged that shows a Kuiper belt with two major components: a dynamically cold population—made of objects on orbits with inclinations $i < 4^\circ$ —and a hot population—made of objects whose inclinations can be as large as 30° , and possibly larger⁹. These populations have different physical properties^{10,11}. In addition, there is a small population (roughly 10% of the total) of objects trapped in mean motion resonances (MMRs) with Neptune. The objects in all three structures have semi-major axes less than ~ 50 AU (refs 12, 13), at which point the Kuiper belt seems to end abruptly (see Fig. 1a).

A study of the literature suggests the following for the early evolution of the Kuiper belt. First, the proto-planetary disk was truncated at roughly 50 AU by one of several possible mechanisms^{14–17}. Then, the dynamical evolution of objects during Neptune's outward migration^{7,18} gave rise to the resonant populations^{6,7}, the hot population^{6,19}, and a related structure known as the scattered disk^{20,21}. This would imply that the objects in the cold population are primordial, presenting us with the mass depletion problem mentioned above.

Here we argue that the initial proto-planetary disk was truncated somewhere near the current location of Neptune (at 30 AU). Placing the edge of the disk at this location provides a natural explanation for why Neptune is where it is—it migrated until it hit the edge of the disk²². Not only the hot population⁶ but also the cold population formed interior to this edge, and was pushed out to its current position during Neptune's migration. This circumvents the mass depletion problem because the mass required for the objects to grow was never in the Kuiper belt. However, the push-out mechanism for the cold population must be different from (but work in conjunction with) that proposed⁶ for the origin of the hot population. This is because that model⁶ involves objects that were scattered by Neptune and thus had their inclinations excited, whereas the inclinations of objects in the cold belt are small. We present a new dynamical mechanism that meets this constraint.

We started our search for a new push-out mechanism by focusing on Neptune's 1:2 MMR. The visible outer edge of the cold belt is close to this MMR's current location at 48 AU (Fig. 1a), which suggested that this resonance might be responsible for this population.

Previous studies⁷ have shown that as Neptune migrated outward, some objects in primordial, low-inclination, low-eccentricity orbits were trapped in this MMR and were pushed outward, along with the resonance. In addition, it was shown that 1:2 MMR does not excite inclinations. Thus, we reasoned that this resonance could have been responsible for the delivery of objects to the region between 40 and 48 AU while keeping the population dynamically cold. However, the standard 'adiabatic model' of resonance trapping predicts that, as the bodies are pushed outwards, their eccentricities monotonically increase. So, as the MMR is pushed through the cold belt, the eccentricities of trapped objects would be larger than those of the observed cold Kuiper-belt objects (see legend of Fig. 1b for details).

But does the adiabatic theory apply to this case? To answer this question, we have simulated, using direct N -body techniques, the evolution of a system containing the four giant planets embedded in a planetesimal disk. The disk extended originally from 10 to 35 AU and contained 50 Earth masses, with a surface density proportional to the inverse heliocentric distance. The disk was modelled using 15,634 particles of equal mass, which interacted with the planets but not with each other, as is usually done in this kind of simulation. The initial locations of Jupiter, Saturn, Uranus and Neptune were at 5.4, 8.7, 13.8 and 18.1 AU respectively. We followed the evolution of this system with the symplectic integrator known as SyMBA^{23,24}. We found that during the integration, the giant planets migrated, owing to the interaction with the disk particles, as expected. In addition,

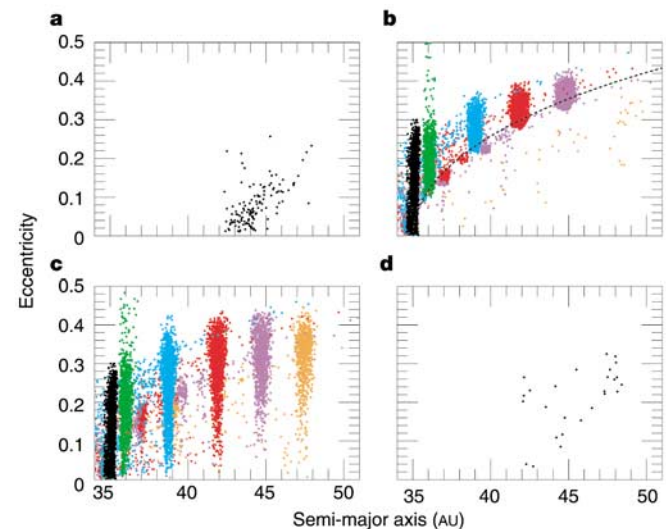


Figure 1 The semi-major axis versus eccentricity distribution for various 'cold' populations. In all cases only objects with inclinations less than 4° are plotted. **a**, Real Kuiper-belt objects. Only those objects that have been observed over multiple oppositions are plotted, to ensure accurate orbits. **b**, Migration if the resonant particles (but not other particles if they are present) are massless. The colours represent different times and thus the 1:2 MMR was at different locations: 6 Myr (black), 8 Myr (green), 16 Myr (blue), 25 Myr (red), 38 Myr (purple) and 98 Myr (orange). At these times, Neptune was at approximately 22 AU, 24 AU, 26 AU, 28 AU and 30 AU, respectively. The black and green data were taken from our direct N -body simulation while the rest were taken from our 'smooth migration' runs. Secondary clumps of a given colour are objects that fell out of the 1:2 MMR and were subsequently trapped in other resonances. The dotted curve is the expected lower limit of the eccentricity as predicted by the adiabatic theory: $\sqrt{1/2 \log(a/a_{\text{edge}})}$, where a is their current semi-major axis and a_{edge} is the primordial outer edge of the disk. So, if we place the original edge at, say, 30 AU, the expected eccentricity at 40 AU should be greater than 0.38. If this were true, we could not produce the cold population, which is between ~ 40 AU and 48 AU and has eccentricities less than ~ 0.2 , with the adiabatic model. **c**, Our simulation in which the resonant particles are massive. The colours have the same meaning as in panel **b**. **d**, Our final simulation with massive disk particles and jumpy migration. The data were taken at 1 Gyr and only those particles with semi-major axes greater than 41 AU were plotted so as to keep it consistent with the definition of the cold belt.

there was indeed a population of low-inclination objects captured in Neptune's 1:2 resonance and pushed beyond the original outer edge of the disk, as predicted by the adiabatic theory.

In contrast with the theory, however, we found that some objects in the MMR have eccentricities that oscillate (Fig. 2) and thus there was always a population of bodies with small eccentricities (see Fig. 1c). These oscillations are the result of a previously unknown secular resonance that is embedded in Neptune's 1:2 MMR (Fig. 3). This secular resonance is generated when there is a significant amount of mass (~ 3 Earth masses) in the 1:2 MMR (see legend of Fig. 3).

How far can the bodies be pushed within the 1:2 MMR and still exhibit large eccentricity oscillations? Unfortunately, we cannot answer this with our current simulations for computational reasons. Because the amount of processing time required for our simulations is linear with respect to the number of disk particles, we were forced to keep the number of particles less than about 10^4 . In reality, the disk was made of many billions of much less massive objects. Consequently, Neptune's migration was much more jumpy in our simulation than it was in reality, which had the effect of artificially clearing objects from the 1:2 MMR (and other resonances) when it was at 38 AU, well before it entered the region of the cold Kuiper belt.

To overcome this problem, we performed a second simulation, in which planetary migration is not directly the result of the interaction with the disk particles, but instead is primarily induced by the inclusion of an artificial drag term in the planets' equations of motion. For this drag we employ the prescription in ref. 7, so that planetary migration is perfectly smooth and exponentially decreasing on a timescale of 40 Myr. The initial orbital configurations of the planets and particles in the 1:2 MMR are those obtained from the previous simulation when the resonance is exactly at the disk's edge. All other disk particles are discarded from the system. The initial total mass of the population in the 1:2 MMR is about three Earth masses. During the integration the disk particles responded to the planets, but not to each other. The planets responded to one another, the disk particles, and the drag force. We found that, throughout the migration process, there were always a number of objects in the 1:2 resonance that had small eccentricities (Fig. 1c). Investigations of the behaviour of these particles show that they are

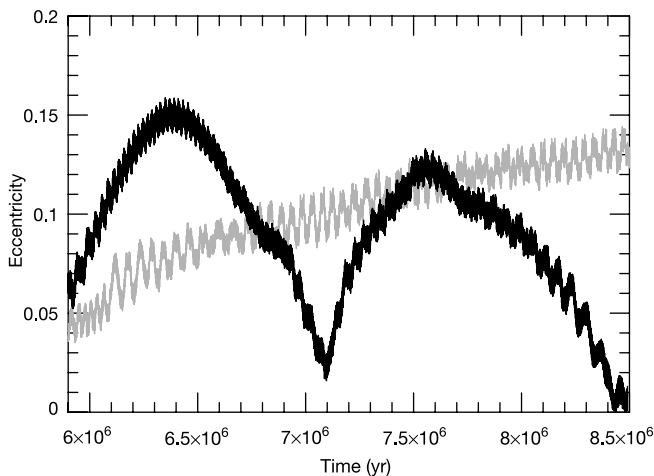


Figure 2 The evolution of eccentricity of particles in the 1:2 MMR in various physical situations. The black curve shows the eccentricity evolution of a resonant particle in our N -body integration after the resonance was pushed beyond the original disk edge (which occurred at $t = 5.9$ Myr). Large-amplitude secular oscillations temporarily drive the eccentricity down to $e \approx 0$. The grey curve shows the behaviour of a resonant particle's eccentricity when the population of massive 1:2 resonant bodies are replaced by an equal population of massless particles. In this case the overall eccentricity evolution is dominated by the monotonic growth, as predicted by the adiabatic theory. Therefore, the mass of the resonant population is the key to having a population of small-eccentricity objects as the 1:2 MMR sweeps through the region now occupied by the cold Kuiper belt.

indeed undergoing oscillations in eccentricity caused by the self-induced secular resonance described above.

So we were now able to produce resonant objects on low eccentricities as the resonance moved between 40 and 48 AU, but we still had not reproduced the observations. At the end of the simulation described in the previous paragraph, most of the objects are in the 1:2 resonance and not in the cold non-resonant population. Our model did not release enough of the objects from the resonance during the migration. This is because these simulations are the opposite extreme of our N -body simulations, that is, the migration is perfectly smooth. Perfectly smooth migration can only occur in an idealized disk composed of an infinite number of infinitely small objects.

In reality, Neptune's migration, and thus that of its MMRs, must have been somewhat jumpy because of the presence of relatively massive objects in the disk²⁵. Thus, we performed another simulation in which we modelled this by providing, at random times, instantaneous 'kicks' to Neptune's velocity in random directions. Figure 1d shows the resulting orbital distribution of the dynamically cold population produced in the simulation in which velocity kicks with magnitude $10^{-4}/\sqrt{p}$ km s⁻¹—where p was a uniform random number between 0 and 1—are applied every 5,000 years. The magnitude and frequency of these kicks statistically correspond to what Neptune would experience if there were one or two objects of about two lunar masses crossing its orbit at any one time. The $1/\sqrt{p}$ dependence is introduced to reproduce the statistical distribution of

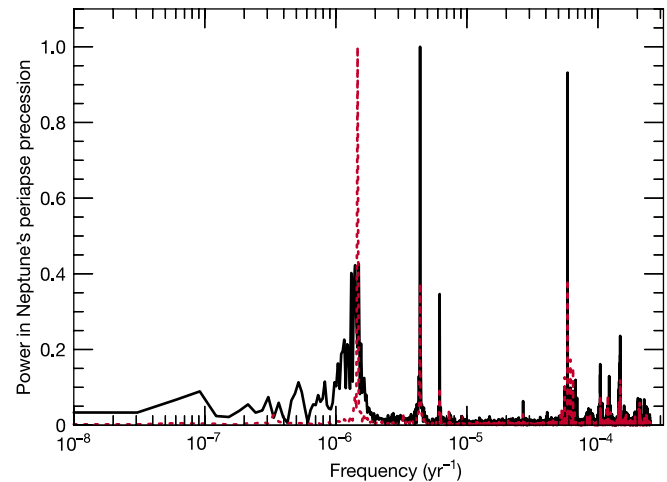


Figure 3 The power spectrum of the periape precession of Neptune's orbit that illustrates the origin of our new secular resonance. A secular resonance occurs when the body's longitude of perihelion precesses at the same rate as one of the frequencies of the planets. In a system consisting of only the four giant planets, the precession of Neptune's perihelion has four independent frequencies. The dotted red curve shows the frequency power spectrum of Neptune's longitude of perihelion in a planetary system where the semi-major axes of Jupiter, Saturn, Uranus and Neptune were at 5.5, 9.3, 15.4 and 22.0 AU, respectively. The slowest frequency is $\sim 1.5 \times 10^{-6}$ yr⁻¹, which is much faster than the precession frequencies of the perihelia of the bodies in the 1:2 MMR. Thus, no secular resonance occurs. However, when a non-negligible (about three Earth masses) amount of mass is placed in Neptune's 1:2 MMR, more frequencies are introduced into the power spectrum of Neptune's perihelion. If the mass were concentrated in one body only one new independent frequency would be introduced. But, if the mass is distributed among a population of objects with a variety of orbits and precession rates, Neptune's precession contains many more frequencies and many of these frequencies overlap, forming bands in the power spectrum. The solid black curve shows the case when we placed 780 particles, with a total of three Earth masses, in the resonance. Note that in this case the spectrum of Neptune exhibits frequencies of roughly 10^{-7} yr⁻¹, which are of the same order as the precession frequencies of the 1:2 MMR particles. Therefore, each body in the 1:2 resonance 'sees' a frequency in Neptune's motion that is close to its own, thereby causing a secular resonance. This situation forces large amplitude oscillations of the eccentricity, like those illustrated in Fig. 2.

impact parameters. There is reasonably good agreement between the observed orbital distribution of the cold population (Fig. 1a) and the results of our simulation. We also note that our new mechanism makes mainly cold objects—63% of the particles between 42 and 50 AU have inclinations less than 4° and all have inclinations less than $\sim 15^\circ$.

We actually performed 12 ‘jumpy migration’ simulations with different values for the magnitude and frequency of the kicks. The run shown in Fig. 1d produced the semi-major axis versus eccentricity distribution most similar to that observed. Reducing the magnitude or the frequency of the kicks leaves progressively more particles in the 1:2 resonance at the end of the simulation, eventually conflicting with the observations. Conversely, increasing the magnitude or the frequency of the kicks leads to simulations where the resonance is depleted significantly before it reaches its final location and thus leaves no objects in the outer Kuiper belt, which again is in conflict with the observations.

The final cold belt population generated by our model represents only $\sim 1\%$ of the population originally in the massive planetesimal disk beyond the initial location of the 1:2 MMR. Obviously, this fraction is model-dependent, but it is necessarily small and explains the small mass of the current population. It is due to a series of successive diminutions related to: (1) the capture efficiency in the 1:2 MMR; (2) the probability of release from the resonance due to Neptune’s encounters with the massive bodies; and (3) the probability of having, at the time of the release from the resonance, an eccentricity sufficiently small to ensure survivability during the subsequent evolutionary phases.

The results presented in this Letter have profound implications for our understanding of the origin of the outer Solar System. In combination with the results in refs 6 and 7, this work implies that the entire Kuiper belt formed well within its current location, and achieved its present-day position as a result of Neptune’s outward migration. Unfortunately, the astronomical community does not, as yet, possess the computational ability to perform simulations of the entire process that include all three mechanisms. Therefore, whether this scheme can explain the ratio of the number of objects in the respective populations (resonant, cold and hot), the exact orbital elements of the Kuiper-belt objects, and the origin of the physical differences between the populations, must remain for future models to address.

However, even in its current form, our scheme does supply a natural explanation for the location of the Kuiper belt’s edge. Although several mechanisms have been proposed to explain the truncation of the proto-planetary disk^{14–17} (a requirement of our model), none can explain why it is near the location of Neptune’s 1:2 MMR. Our model explains this observation as a natural consequence of the dynamical evolution after the edge-forming event. Indeed, it predicts that although a few objects may eventually be discovered beyond the resonance on low-inclination, nearly circular orbits (because they can be formed by other dynamical mechanisms), they will be rare. Hence the Kuiper belt’s effective edge should be precisely at the current location of Neptune’s 1:2 MMR. □

Received 19 April; accepted 14 October 2003; doi:10.1038/nature02120.

1. Trujillo, C. A., Jewitt, D. C. & Luu, J. X. Properties of the Trans-Neptunian belt: statistics from the Canada-France-Hawaii telescope survey. *Astron. J.* **122**, 457–473 (2001).
2. Gladman, B. *et al.* The structure of the Kuiper belt: Size distribution and radial extent. *Astron. J.* **122**, 1051–1066 (2001).
3. Stern, S. A. & Colwell, J. E. Accretion in the Edgeworth-Kuiper belt: forming 100–1000 km radius bodies at 30 AU and beyond. *Astron. J.* **114**, 841–849 (1997).
4. Kenyon, S. J. & Luu, J. X. Accretion in the early Kuiper belt. I. Coagulation and velocity evolution. *Astron. J.* **115**, 2136–2160 (1998).
5. Morbidelli, A. & Brown, M. E. The Kuiper belt and the primordial evolution of the solar system. In *Comets II* (eds Festou, M., Keller, H. U. & Weaver, H.) (Univ. Arizona Press, Tucson, in the press).
6. Gomes, R. S. The origin of the Kuiper belt high inclination population. *Icarus* **161**, 404–418 (2003).
7. Malhotra, R. The origin of Pluto’s orbit: implications for the Solar System beyond Neptune. *Astron. J.* **110**, 420–432 (1995).
8. Jewitt, D. C. & Luu, J. X. Discovery of the candidate Kuiper belt object 1992 QB1. *Nature* **362**, 730–732 (1993).

9. Brown, M. The inclination distribution of the Kuiper belt. *Astron. J.* **121**, 2804–2814 (2001).
10. Levison, H. F. & Stern, S. A. On the size dependence of the inclination distribution of the main Kuiper belt. *Astron. J.* **121**, 1730–1735 (2001).
11. Trujillo, C. A. & Brown, M. E. A correlation between inclination and color in the classical Kuiper belt. *Astrophys. J.* **566**, 125–128 (2002).
12. Trujillo, C. A. & Brown, M. E. The radial distribution of the Kuiper belt. *Astrophys. J.* **554**, 95–98 (2001).
13. Allen, R. L., Bernstein, G. M. & Malhotra, R. Observational limits on a distant cold Kuiper belt. *Astron. J.* **124**, 2949–2954 (2002).
14. Ida, S., Larwood, J. & Burkert, A. Evidence for early stellar encounters in the orbital distribution of Edgeworth-Kuiper belt objects. *Astrophys. J.* **528**, 351–356 (2000).
15. Weidenschilling, S. Formation of planetesimals/cometesimals in the solar nebula. In *Comets II* (eds Festou, M., Keller, H. U. & Weaver, H.) (Univ. Arizona Press, Tucson, in the press).
16. Hollenbach, D. & Adams, F. C. Dispersal of disks around young stars: constraints on Kuiper belt formation. In *Debris Disks and the Formation of Planets* (eds Caroff, L. & Backman, D.) (ASP, San Francisco, in the press).
17. Stone, J. M., Gammie, C. F., Balbus, S. A. & Hawley, J. F. in *Protostars and Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 589–612 (Univ. Arizona Press, Tucson, 1998).
18. Fernández, J. A. & Ip, W. H. Orbital expansion and resonant trapping during the late accretion stages of the outer planets. *Planet. Space Sci.* **44**, 431–439 (1996).
19. Morbidelli, A. & Levison, H. F. Kuiper belt interlopers. *Nature* **422**, 30–31 (2003).
20. Duncan, M. & Levison, H. A scattered disk of icy objects and the origin of Jupiter-family comets. *Science* **276**, 1670–1672 (1997).
21. Luu, J. *et al.* A new dynamical class in the trans-Neptunian Solar System. *Nature* **387**, 573 (1997).
22. Gomes, R. S., Morbidelli, A. & Levison, H. F. Planetary migration in a planetesimal disk: why did Neptune stop at 30 AU? *Icarus* (submitted).
23. Duncan, M. J., Levison, H. F. & Lee, M.-H. A multiple timestep symplectic algorithm for integrating close encounters. *Astron. J.* **116**, 2067–2077 (1998).
24. Wisdom, J. & Holman, M. Symplectic maps for the n-body problem. *Astron. J.* **102**, 1528–1538 (1991).
25. Hahn, J. M. & Malhotra, R. Orbital evolution of planets embedded in a planetesimal disk. *Astron. J.* **117**, 3041–3053 (1999).

Acknowledgements We are grateful to R. Gomes and M. Holman for acting as referees on this paper. We also thank L. Dones, W. Bottke, J.-M. Petit, and K. Tsiganis for help with an early version of the text and M. Duncan and R. Gomes for discussions. H.F.L. is grateful for funding from NASA. We thank the CNRS and NSF for encouraging friendly relationships between the US and France.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be sent to H.F.L. (hal@boulder.swri.edu).

Microfluidic sorting in an optical lattice

M. P. MacDonald¹, G. C. Spalding^{1,2} & K. Dholakia¹

¹School of Physics and Astronomy, University of St Andrews, North Haugh, St Andrews, Fife KY16 9SS, UK

²Department of Physics, Illinois Wesleyan University, Bloomington, Illinois 61702, USA

The response of a microscopic dielectric object to an applied light field can profoundly affect its kinetic motion¹. A classic example of this is an optical trap, which can hold a particle in a tightly focused light beam². Optical fields can also be used to arrange, guide or deflect particles in appropriate light-field geometries^{3,4}. Here we demonstrate an optical sorter for microscopic particles that exploits the interaction of particles—biological or otherwise—with an extended, interlinked, dynamically reconfigurable, three-dimensional optical lattice. The strength of this interaction with the lattice sites depends on the optical polarizability of the particles, giving tunable selection criteria. We demonstrate both sorting by size (of protein microcapsule drug delivery agents) and sorting by refractive index (of other colloidal particle streams). The sorting efficiency of this method approaches 100%, with values of 96% or more observed even for concentrated solutions with throughputs exceeding those reported for fluorescence-activated cell sorting⁵. This powerful, non-invasive technique is suited to sorting and fractionation within integrated

('lab-on-a-chip') microfluidic systems, and can be applied in colloidal, molecular and biological research.

In microfluidics, flow is predominantly laminar, creating challenges in the design of actuators such as mixers and sorters. The extremely low Reynolds numbers associated with microfluidic flows (for example, 10^{-3}) mean that inertial effects cannot be used to sort particles, yet the ability to select and sort both colloidal and biological matter—in a fast and efficient manner related to its physical properties—is a key requirement in this environment. Existing techniques^{6,7} for separation of macro-molecules typically require batch processing, while the most novel alternatives are reliant upon diffusion^{8–10}. Such approaches cannot separate objects as large as cells or other colloidal suspensions. The current standard for separating such suspensions, fluorescence-activated cell sorting (FACS), requires fluorescence to distinguish between cell types¹¹. For particles that do not fluoresce, sedimentation¹² can be used but this gives low resolution and is not suitable for small volumes of analyte. Where small volumes are involved the trend is towards integrating devices into micro-total analysis systems, or 'lab-on-a-chip'⁵. Here we introduce an effective, all-optical, dynamically reconfigurable means of sorting in a microfluidic sample cell. This powerful yet simple technique is non-invasive, sorting particles without physical contact. Sole use of optical forces simplifies surface interaction and sterility issues by removing the extremely high surface area associated with any physical sieve or gel.

Recent research into sculpting optical wavefronts and intensity profiles has led to enhanced levels of control over microscopic matter and biological particles¹³. Interferometric patterns of light can create one-, two- or three-dimensional optical lattices, enabling numerous studies in atomic and colloidal science^{4,13,14}. Here we exploit the influence of optical lattices on matter to enforce the strong lateral separation required for continuous processing while maintaining compatibility with the full mesoscopic range of particle sizes (up to the scale of $\sim 100\ \mu\text{m}$). The interaction between optical lattices and matter depends upon a relative polarizability that grows with the size of the particle yet remains effective down to atomic scales¹⁵. Selected particle types follow defined paths through an optical lattice, providing lateral fractionation. Importantly, the optical lattice is three-dimensional in nature, providing the ability to sort particles throughout a three-dimensional flow.

A two-dimensional array of traps has been used recently¹⁶ to explore the incursion—from all in-plane angles simultaneously—of colloid into the optical array. This led to a study⁴ of the angular deviation of single, well-isolated (that is, extremely dilute) particle trajectories as a function of a relative rotation between an externally imposed fluid flow and the [100] direction of the trap array. These studies used solely monodisperse colloidal solutions, and thus did not demonstrate fractionation or separation of mixed particle streams, but there was a suggestion that optical arrays might be used for that purpose⁴. A key observation was that this array of traps could provide only weak angular deflection (no more than 11°). Recognizing the limitation of using discrete (point-like), isotropic traps, we instead utilize an extended, three-dimensional optical lattice of interlinked sites, with linkages established along the desired channelling direction. We show that the optimum lattice type for high-angle separation lies between the two extremes of discrete traps and extended guides, where we have strongly interlinked localized intensity maxima.

Here a five-beam interference pattern is used to sort co-streaming particles in microfluidic flows. Figure 1 describes the overall concept for a sorting machine based on optical forces—interaction with optical fields provides a selective means of removing material matching specific criteria from an otherwise laminar stream. When a flow of mixed particles is passed through the lattice, selected particles are strongly deflected from their original trajectories while others pass straight through largely unhindered, depending upon their sensitivity to the optical potential. In our apparatus, a single

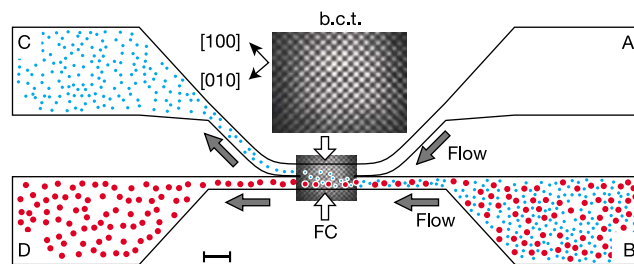


Figure 1 The concept of optical fractionation. Low Reynolds number flows will be laminar: without an actuator all particles from chamber B would flow into chamber D. Chamber A would typically introduce a 'blank' flow stream, although this could be any stream into which the selected particles are to be introduced. By introducing a three-dimensional optical lattice—in this case a body-centred tetragonal (b.c.t.) lattice—into the fractionation chamber (FC), one species of particle is selectively pushed into the upper flow field. The reconfigurability of the optical lattice allows for dynamic updating of selection criteria. For weakly segregated species, the analyte can be either recirculated through the optical lattice or directed through cascaded separation chambers. This latter option also allows the use of multiple selection criteria in a single integrated chip. The flow volume in our current sample cells is $100\ \mu\text{m}$ thick; scale bar, $40\ \mu\text{m}$.

$1,070\text{-nm}$ laser beam (coherence length $1.2\ \text{cm}$) passes through a diffractive beam splitter, producing four beams diverging from the central, undiffracted spot at 4.6° , in a cross shape (Supplementary Fig. A). Collimating optics provide an 'infinity space' in which we have independent control over the phase and amplitude of each of these five beams before co-focusing through an aspheric lens produces a large, three-dimensional optical lattice through multi-beam interference.

First we demonstrate purely chemical separation (Fig. 2); a co-streaming mixture of same-sized ($2\ \mu\text{m}$ diameter) silica and polymer spheres passes through our optical lattice. There is 45° angular separation—even at flow speeds in excess of $35\ \mu\text{m s}^{-1}$ (for a laser output of $530\ \text{mW}$ and densities as shown). Differences in contrast allow the two, co-flowing species to be separately tracked, via automated particle identification with tenth-pixel resolution¹⁷ (here $\sim 8\ \text{nm}$). We have used this precision to assemble a variety of statistics (such as the size and location distributions of particle displacements¹⁸) for a number of polydisperse systems passing through the lattice.

One measure of efficiency, appropriate for two-species flows (Fig. 2), is an outcome-based figure of merit (FOM) given by the percentage of species A deflected by the optical lattice minus the percentage of species B deflected. Figure 3 shows the experimental results for this efficiency FOM, as a function of flow speed, taken from statistical analysis of more than 1,200 particle trajectories in a dense flow of mixed $2\text{-}\mu\text{m}$ polymer and silica spheres. As the flow speed increases, we see changes in the particle displacement statistics¹⁸, from an exponential distribution at low speeds (characteristic of hopping transport) to a gaussian distribution (reflecting guided flow, which can be further tuned from pseudo-ballistic to diffusively guided regimes and—finally—to the limit where all particles flow past unhindered). Between the regimes of hopping and free flow, we observe selective dynamics¹⁸ and, accordingly, the value of the FOM rises, approaching 100% in the semi-dilute limit, with values of 96% or more sustained even for concentrated solutions as shown in Fig. 3 (an earlier proposal to use only a single laser beam¹⁹ achieved only a fraction of one per cent in efficiency). Three mechanisms account for reductions from 100% efficiency: if the polydisperse flow is not well covered by the optical lattice, if the flow is beyond the semi-dilute limit (particularly if large aggregates are present), or if the flow rate is not optimized for the laser power (which amounts to a balance between kinetic and optical potential energies¹⁸). The influence of the first two of these mechanisms is kept to a minimum

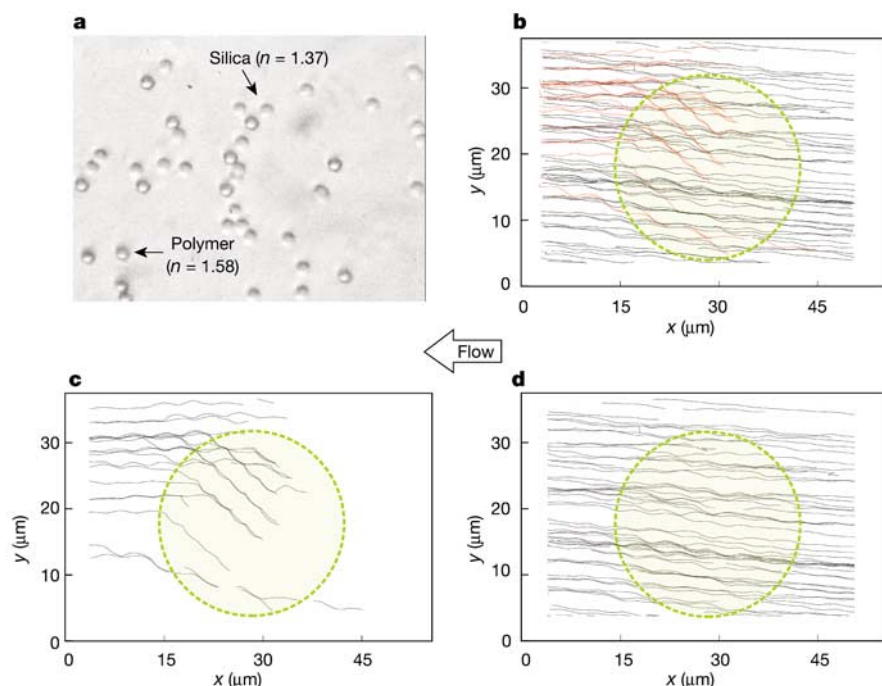


Figure 2 Separation by index of refraction. Same-sized ($2\text{ }\mu\text{m}$ diameter) polymer and silica spheres in water co-flow from right to left through the optical lattice at a fluid speed of $30\text{ }\mu\text{m s}^{-1}$ with a total incident laser power of 530 mW. **a**, An image of the silica/polymer mix, indicating the typical particle density. n , index of refraction. **b**, Polymer trajectories are shown in red, and silica trajectories in black, with a green circle indicating

the x - y range over which the optical lattice is most intense. **c**, The polymer tracks show a deflection in excess of 45° while **(d)** the silica tracks are only slightly modulated by the optical lattice. This very large angular separation is attributed to the linking of intensity maxima along the 100 direction of the optical lattice.

by the use of high-angle deflections in the optical lattice.

Each of these limitations upon efficiency can be rephrased as a limitation upon the overall throughput of the system. Typical throughput of our prototypical system was ~ 25 particles per second, slightly higher than those reported for separation in a microfabricated FACS. Optical fractionation does not require tagging or endogenous fluorophores. Additionally, optical fractionation will allow highly parallel sorting, with throughputs that will scale with the volume of the sample cell, providing high parallel throughput and $\sim 100\%$ efficiency for deflection of, for example, a selected cell type, with complete exclusion of other types.

Using the set-up described above, and 530 mW of power (measured directly at the laser), a monodisperse ($2\text{ }\mu\text{m}$) collection of protein microcapsules is extracted from a polydisperse particle stream moving at $20\text{ }\mu\text{m s}^{-1}$ (Fig. 4). These microcapsules are an ultrasound contrast agent that can be used to locally permeate cell membranes (a technique known as sonoporation²⁰), providing direct DNA transfection or drug delivery. Using optical fractionation to create a collection of protein microcapsules of a tailored size provides an essential level of control for studies of drug delivery at the cellular level. Practical biological potential (specifically for haematology or immunology) has also been demonstrated by separating erythrocytes (red blood cells) from lymphocytes dispersed in a culture medium (M.P.M., L. Paterson, G.C.S., A. Richies and K.D., manuscript in preparation).

Whereas other technologies have trouble distinguishing non-fluorescent particle sizes that differ by just 20%, optical fractionation offers unparalleled exponential size selectivity: the probability for activation over a potential barrier involves a Boltzmann factor, an exponential dependence upon the ratio of (1) the additional energy required to overcome the barrier to (2) the relevant thermal energy scale, which is set by the temperature. Term (1) is dependent upon particle size, resulting in an exponential sensitivity to size. This result is robustly general, even as we tune the interconnectivity

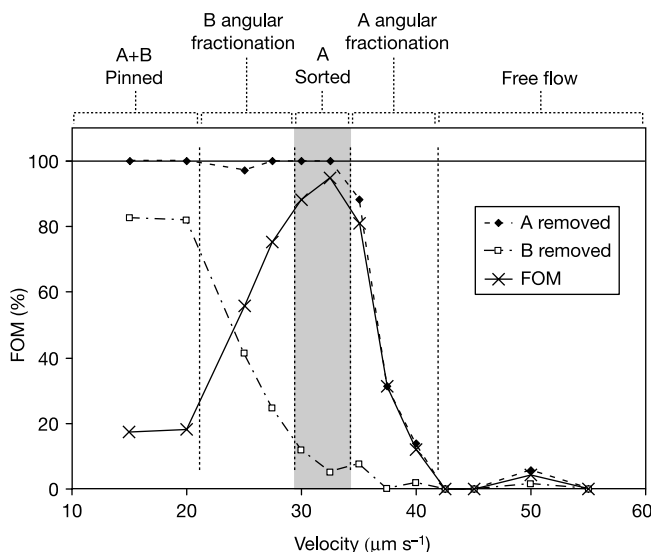


Figure 3 Typical efficiencies. Our efficiency figure of merit (FOM) is, for the co-flowing mixtures of two species shown in Fig. 2, the percentage of species A (polymer) deflected minus the percentage of species B (silica) deflected, plotted as flow speed increases, or laser power is reduced. At speeds just beyond de-pinning, both species are deflected by the optical lattice. Appropriate choice of particle speeds and laser power yields optimal separation. The data shown are based on more than 1,200 particle trajectories, with a laser output of 530 mW. For each material system, similar statistics are gathered in order to establish acceptable operating parameters (here, for example, the optimal measured flow speed was $32.5\text{ }\mu\text{m s}^{-1}$). The particles were suspended in water. Use of a more viscous fluid such as a culture medium simply reduces the optimum flow velocity for optical sorting, but does not change the general behaviour of the technique.

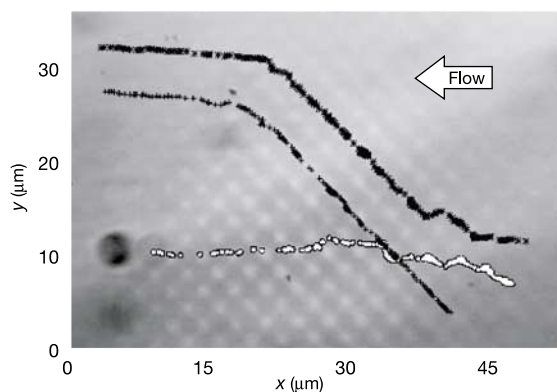


Figure 4 Optical fractionation by size. Black crosses represent the frame-by-frame positions (tracked at standard video rates with ~ 8 nm precision) of two $2\text{-}\mu\text{m}$ diameter protein microcapsules (drug delivery agents) as they flow from right to left across the optical lattice. The flow speed is $20\text{ }\mu\text{m s}^{-1}$ with a total incident laser power of 530 mW. Again, significant angular deflection is achieved, while a co-flowing $4\text{-}\mu\text{m}$ -diameter capsule of the same sort flows nearly straight through (white dots). The collection of $2\text{-}\mu\text{m}$ capsules allows a controlled study of drug delivery at the cellular level.

of the lattice. Such anisotropy merely changes a prefactor in term (1), for motion along the direction of lowered barrier heights.

We note that it is necessary to move beyond the limit of discrete, isotropic traps to achieve the sorts of concentrated, mixed-flow separations demonstrated here. Rather than using kinetic lock-in of hopping probabilities⁴, we provide a more deterministic form of channelling. We use light patterns that contain a tailored degree of 'interconnectedness' (that is, lower potential energy barriers) along the direction of desired channelling¹. Supplementary Fig. B provides details of the experimental parameters used to tune these linkages, along with detailed mapping of lattice parameters and the linkages themselves. The details of the linkages between lattice sites plays a key role in optical fractionation. Distinct individual lattice points are ideally suited to trapping particles, whereas interlinks between the lattice sites are central to deflecting particles into a desired trajectory. The optimum lattice for fractionation with high angles of displacement lies between a discrete lattice of intensity maxima and continuous guides; we went on to find that the body-centred tetragonal (b.c.t.) lattice, used for the results shown in Figs 2–4, is satisfactory. This is demonstrated clearly in Supplementary Fig. D, where the linked b.c.t. lattice has both the highest efficiency and the highest throughput when fractionating at 45° . More detailed calculations and discussion of this will appear elsewhere¹⁸.

Arrays composed of discrete, isotropic traps tend to 'jam' (which is why studies involving such arrays tend to use extremely dilute flows). The dwell time in strongly localizing traps significantly slows particles with respect to the input stream. Jamming is likely to occur in long arrays of this type, effectively preventing particles that would otherwise pass through un-deflected from doing so, and leading to 'spill-over' contamination of the uptake stream.

Practical fractionating arrays must be long enough to transport the selected particles (transversely) from the most distant portion of the input stream all the way to the uptake stream, where they are extracted. For dense flows, greater length is required, but in general the larger the angle of deflection through the lattice, the shorter the lattice can be, making fractionation at 45° with a linked b.c.t. lattice particularly well suited to practical implementation of optical fractionation. □

Received 1 July; accepted 21 October 2003; doi:10.1038/nature02144.

1. Tatarikova, S. A., Sibbett, W. & Dholakia, K. Brownian particle in an optical potential of the washboard type. *Phys. Rev. Lett.* **91**, 038101 (2003).
2. Ashkin, A., Dziedzic, J. M., Bjorkholm, J. E. & Chu, S. Observation of a single-beam gradient force optical trap for dielectric particles. *Opt. Lett.* **11**, 288–290 (1986).

3. Burns, M. M., Fournier, J. M. & Golovchenko, J. A. Optical matter—crystallization and binding in intense optical-fields. *Science* **249**, 749–754 (1990).
4. Korda, P. T., Taylor, M. B. & Grier, D. G. Kinetically locked-in colloidal transport in an array of optical tweezers. *Phys. Rev. Lett.* **89**, 128301 (2002).
5. Fu, A. Y., Spence, C., Scherer, A., Arnold, F. H. & Quake, S. R. A microfabricated fluorescence-activated cell sorter. *Nature Biotechnol.* **17**, 1109–1111 (1999).
6. Han, J. & Craighead, H. G. Separation of long DNA molecules in a microfabricated entropic trap array. *Science* **288**, 1026–1029 (2000).
7. Nykypanchuk, D., Strey, H. H. & Hoagland, D. A. Brownian motion of DNA confined within a two-dimensional array. *Science* **297**, 987–990 (2002).
8. Ertas, D. Lateral separation of macromolecules and polyelectrolytes in microlithographic arrays. *Phys. Rev. Lett.* **80**, 1548–1551 (1998).
9. Duke, T. A. J. & Austin, R. H. Microfabricated sieve for the continuous sorting of macromolecules. *Phys. Rev. Lett.* **80**, 1552–1555 (1998).
10. Chou, C. F. et al. Electroless dielectrophoresis of single- and double-stranded DNA. *Biophys. J.* **83**, 2170–2179 (2002).
11. Galbraith, D. W., Anderson, M. T. & Herzenberg, L. A. in *Methods in Cell Biology* Vol. 58 (eds Sullivan, K. F. & Kay, S. A.) 315–341 (Academic, London, 1999).
12. Athanasopoulou, A., Koliadima, A. & Karaiskakis, G. New methodologies of field-flow fractionation for the separation and characterization of dilute colloidal samples. *Instrum. Sci. Technol.* **24**, 79–94 (1996).
13. Dholakia, K., Spalding, G. C. & MacDonald, M. Optical tweezers: The next generation. *Phys. World* **15**, 31–35 (2002).
14. MacDonald, M. P. et al. Creation and manipulation of three-dimensional optically trapped structures. *Science* **296**, 1101–1103 (2002).
15. Greiner, M., Mandel, O., Esslinger, T., Hansch, T. W. & Bloch, I. Quantum phase transition from a superfluid to a Mott insulator in a gas of ultracold atoms. *Nature* **415**, 39–44 (2002).
16. Korda, P. T., Spalding, G. C. & Grier, D. G. Evolution of a colloidal critical state in an optical pinning potential landscape. *Phys. Rev. B* **66**, 024504 (2002).
17. Crocker, J. C. & Grier, D. G. Methods of digital video microscopy for colloidal studies. *J. Colloid Interface Sci.* **179**, 298–310 (1996).
18. MacDonald, M. P., Spalding, G. C. & Dholakia, K. Transport and fractionation of brownian particles in an optical lattice. *Phys. Rev. Lett.* (submitted).
19. Imasaka, T., Kawabata, Y., Kaneta, T. & Ishidzu, Y. Optical chromatography. *Anal. Chem.* **67**, 1763–1765 (1995).
20. Marmottant, P. & Hilgenfeldt, S. Controlled vesicle deformation and lysis by single oscillating bubbles. *Nature* **423**, 153–156 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank P. Campbell for supplying protein microcapsules, and A. Riches for blood samples. This work was supported by the UK Engineering and Physical Sciences Research Council, the Research Corporation, and the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.P.M. (mpm4@st-and.ac.uk).

Low-temperature processing of 'baroplastics' by pressure-induced flow

Juan A. Gonzalez-Leon¹, Metin H. Acar¹, Sang-Woog Ryu¹, Anne-Valérie G. Ruzette^{1*} & Anne M. Mayes¹

¹Department of Materials Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA

* Present address: Laboratoire Matière Molle et Chimie, ESPCI, 10 rue Vauquelin, F-75231 Paris cedex 05, France

The manufacturing of plastics traditionally involves melt processing at temperatures typically greater than 200°C —to enable extrusion or moulding under pressure into desired forms—followed by solidification. This process consumes energy and can cause substantial degradation of polymers and additives (such as flame retardants and ultraviolet stabilizers), limiting plastics performance and recyclability¹. It was recently reported that the application of pressure could induce melt-like behaviour in the block copolymer polystyrene-*block*-poly(*n*-butyl methacrylate) (PS-*b*-PBMA)², and this behaviour has now been demon-

strated in a range of other block copolymer systems^{3–8}. These polymers have been termed baroplastics^{2–5}. However, in each case, the order-to-disorder transition, which gives rise to the accompanying change in rheology from soft solid to melt^{9,10}, was observed at temperatures far exceeding the glass transition temperatures (T_g) of both components. Here we show that baroplastic systems containing nanophase domains of one high- T_g and one low- T_g component can exhibit melt-like flow under pressure at ambient temperature through an apparent semi-solid partial mixing mechanism that substantially preserves the high- T_g phase. These systems were shredded and remoulded ten times with no evident property degradation. Baroplastics with low-temperature formability promise lower energy consumption in manufacture and processing, reduced use of additives, faster production and improved recyclability, and also provide potential alternatives to current thermoplastic elastomers, rubber-modified plastics, and semi-crystalline polymers^{11,12}.

To design materials molecularly, incorporating a low- T_g component that can solvate (or partially solvate) a high- T_g component under pressure, we employed principles for polymer pair miscibility recently elucidated through a straightforward extension of the Flory–Huggins regular solution model for polymer mixtures to compressible systems^{4,13,14}. The free energy per unit volume of a binary compressible polymer mixture can be written as:

$$\Delta g_{\text{mix}} = kT \left[\frac{\phi_A \tilde{\rho}_A}{N_A v_A} \ln \phi_A + \frac{\phi_B \tilde{\rho}_B}{N_B v_B} \ln \phi_B \right] \quad (1)$$

$$+ \phi_A \tilde{\rho}_A \phi_B \tilde{\rho}_B (\delta_{A,0} - \delta_{B,0})^2 + \phi_A \phi_B (\tilde{\rho}_A - \tilde{\rho}_B) (\delta_A^2 - \delta_B^2)$$

where k is the Boltzmann constant; ϕ_i , N_i and v_i are the volume fraction, number of segments per chain, and hard core segmental volume of the i th component; $\delta_{i,0}^2$ and δ_i^2 are the cohesive energy densities at 0 K and temperature T , respectively; and $\tilde{\rho}_i = \rho_i / \rho_i^*$ is the reduced density, with the hard core density defined as $\rho_i^* = M_{u,i} / N_0 v_i$, where $M_{u,i}$ is the monomer molecular mass and N_0 is Avogadro's number. In equation (1), the first two terms are analogous to the Flory–Huggins entropy (–) and enthalpy (+) of mixing terms, while the third term arises solely from compressibility. The third term can either favour (–) or oppose (+) segmental mixing, tending to zero in the incompressible limit. Equation (1) can be used to predict phase diagrams solely from pure component properties, and has successfully captured the qualitative phase behaviour of over 30 weakly interacting polymer pairs^{4,13,14}. In the present context, systems for which the third term approaches zero at 0 K from negative values, and whose ambient-state densities nearly match¹⁵ ($0.94 \rho_A < \rho_B < 1.06 \rho_A$), have been found to exhibit pressure-induced miscibility, including PS/PBMA², PS/poly(hexyl methacrylate)^{3,4}, PS/poly(pentyl methacrylate)⁵, polybutadiene/polyisoprene (PB/PI)⁸ and poly(ethylene propylene)/poly(ethyl methylene) (PEP/PEE)⁹. By contrast, these conditions are not met for PS/PI or PS/PB, two commercially

important block copolymer systems that have been found to exhibit reduced miscibility with applied pressure^{16–19}. Using the above criteria, two new baroplastic candidates with one high- T_g and one low- T_g component were identified, namely, polystyrene/poly(*n*-butyl acrylate) (PS/PBA, $T_{g,\text{PBA}} \approx -54^\circ\text{C}$) and polystyrene/poly(2-ethylhexyl acrylate) (PS/PEHA, $T_{g,\text{PEHA}} \approx -50^\circ\text{C}$). Table 1 lists the molecular and compositional characteristics of the materials discussed in this study.

The ability to process PS-*b*-PBA and PS-*b*-PEHA block copolymers at room temperature despite their high- T_g PS component was demonstrated by compression-moulding the freeze-dried precipitates into rigid transparent objects using a standard hydraulic press and moulds machined from stainless steel or aluminium. Figure 1a shows the starting material and final products of a 38,000 g mol^{–1} PS-*b*-PBA block copolymer incorporating 45 wt% PBA moulded at 25 °C and 34.5 MPa (5,000 p.s.i.) for 5 min. The transparency of the moulded objects and their accuracy of form are testimony that the copolymer flowed under applied pressure to take the shape of its

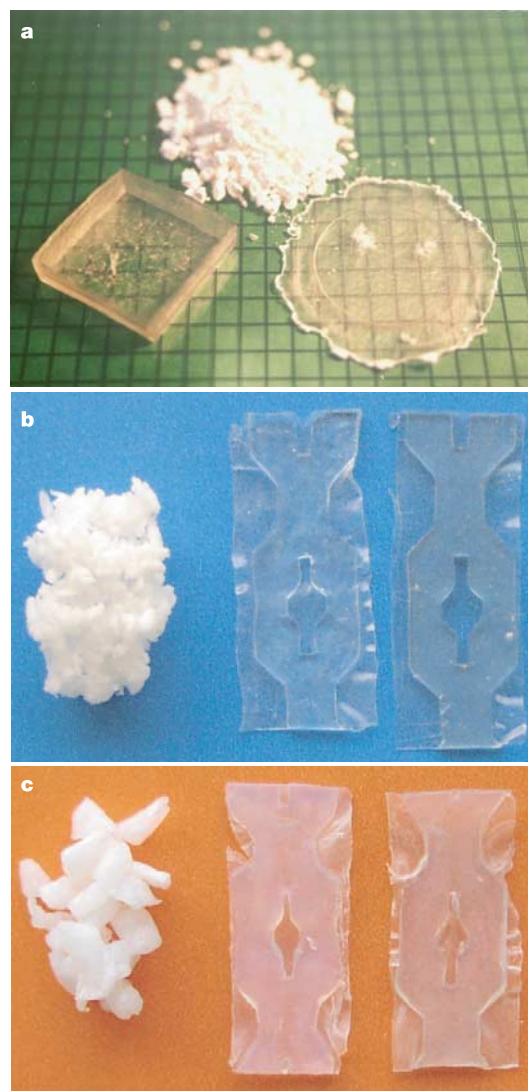


Figure 1 Processed baroplastics samples. **a**, 38,000 g mol^{–1} polystyrene-*block*-poly(*n*-butyl acrylate), PS-*b*-PBA, with 45 wt% PBA as obtained from freeze-drying and after processing by compression moulding at 25 °C using a pressure of 34.5 MPa (5,000 p.s.i.). **b**, 60,000 g mol^{–1} polystyrene-*block*-poly(2-ethylhexyl acrylate), PS-*b*-PEHA, with 52 wt% PEHA reprocessed once and ten times at 30 °C under a pressure of 34.5 MPa. **c**, PEHA/PS core-shell material with 52 wt% PEHA processed one time and recycled ten times at 25 °C with a pressure of 34.5 MPa.

Table 1 Molecular and compositional characteristics of baroplastics

System	PS (wt%)*	Molecular mass, M_n (g mol ^{–1})†	Average size (nm)‡
Block copolymers			
PS- <i>b</i> -PBA	70	100,000	
PS- <i>b</i> -PBA	63	73,000	
PS- <i>b</i> -PBA	55	38,000	
PS- <i>b</i> -PEHA	48	60,000	
Core-shell			
PEHA/PS	50	300,000/200,000	67
PBA/PS- δ_8	66	320,000/120,000	87

*Determined by ¹H-NMR.

†Molecular weight measured by GPC based on PS standard.

‡Particle size determined by dynamic light scattering using a Brookhaven Instruments Co. Zeta Potential Analyzer.

container. For example, the lid of a plastic sample holder box was copied to sufficient accuracy to provide a tight seal with the original box. A $73,000 \text{ g mol}^{-1}$ PS-*b*-PBA sample containing 27 wt% PBA could be processed at 80°C at 34.5 MPa.

The premise that low-temperature processing capability arises from the pressure-enhanced miscibility of the PS and PBA block components is supported by *in situ* small-angle neutron scattering (SANS) data taken at elevated temperatures ($120^\circ\text{C} < T < 200^\circ\text{C}$) on a $100,000 \text{ g mol}^{-1}$ PS-*b*-PBA system with 70 wt% PS (data not shown). A pressure coefficient of $dT_{\text{ODT}}/dP \approx -100^\circ\text{C kbar}^{-1}$ was estimated² for this material, which disorders with increasing temperature (ODT, order to disorder transition).

Moulding studies on a $60,000 \text{ g mol}^{-1}$ PS-*b*-PEHA block copolymer incorporating 52 wt% PEHA similarly revealed low-temperature processing capability, and further illustrated the recycling potential of baroplastic block copolymers. Figure 1b shows moulded samples reprocessed once and ten times at 34.5 MPa and 30°C for 5 min. Between consecutive mouldings, samples were shredded into $\sim 3 \text{ mm}$ pieces to create the feed. Although detailed measurements have not been performed to date, the mechanical and optical properties (discounting particulate inclusions) after ten moulding operations appear comparable to the original moulded specimen.

To probe how low-temperature processing influences morphology, SANS was performed on samples before and after room-temperature processing. Figure 2a shows SANS data for the same PS-*b*-PEHA block copolymer in its initial freeze-dried state and after processing once and ten times at 34.5 MPa and 25°C for 5 min. The sample before moulding shows a sharp reflection at wavevector $q \approx 0.025 \text{ \AA}^{-1}$, denoting the nanoscale separation of the PS and PEHA blocks with a periodicity of 25 nm. After processing, the scattering maximum is observed to broaden and decrease in intensity, suggesting that the PS and PEHA blocks are made at least partially miscible under pressure.

Differential scanning calorimetry (DSC) traces on this same material freeze-dried and after ten recycles at 25°C are shown in Fig. 3a. Two glass transitions are observed in each sample near -36 and 55°C . The fact that the material continues to exhibit a glassy-phase T_g well above the processing temperature of 25°C is an interesting clue to the processing mechanism. One would expect that if the copolymer were driven into the fully disordered state during moulding, upon the removal of pressure, demixing of the blocks would be arrested kinetically, that is, when the glass transition of the forming glassy phase equals the processing temperature. Instead, the glassy-state T_g remains well above the processing temperature—a clear advantage for applications. A likely explanation is that the copolymer does not fully disorder under pressure, but instead retains some glassy regions, processing as a semi-solid.

This view of the baroplastics processing mechanism is supported by experiments on a PS-*b*-PEHA sample processed at 34.5 MPa that was subsequently annealed at 85°C overnight and reprocessed at 34.5 MPa. SANS data for the processed, annealed and reprocessed samples are shown in Fig. 2b. Upon annealing, the scattering maximum intensifies and narrows, indicative of enhanced block demixing and/or domain order. The corresponding DSC trace (Fig. 3b) shows the emergence of a glass transition at $\sim 85^\circ\text{C}$, with the glass transitions at -35 and 55°C still present. After remoulding, the peak in the SANS data reverts back to the shape characteristic of the processed state. The DSC trace, however, shows that this sample retains regions with a $T_g \approx 85^\circ\text{C}$ —that is, 60° above the processing temperature. The fact that these regions are not seen in the initially processed sample, but only after the elevated temperature annealing, provides strong supporting evidence for a semi-solid processing mechanism where the high- T_g phase is retained.

The low-temperature processing of block copolymers is thought to be facilitated by their morphology, which presents a high degree

of interface between the two polymer components. From a commercial viewpoint, however, block copolymers have potential drawbacks as substitutes for today's commodity plastics in that their synthesis is generally more complex and expensive, while the number of polymers that can be readily incorporated into block copolymers is limited^{20–22}. As an alternative approach to the preparation of baroplastics, core-shell nanoparticle systems were investigated. Similar to block copolymer nanophases, core-shell nanoparticles contain domains of one component that are molecularly adjacent to those of the second component. Such systems can be made by emulsion polymerization of sequentially added monomers²³, as performed here, or by emulsification and surface precipitation of previously synthesized polymers, allowing for a substantially wider array of baroplastics candidates.

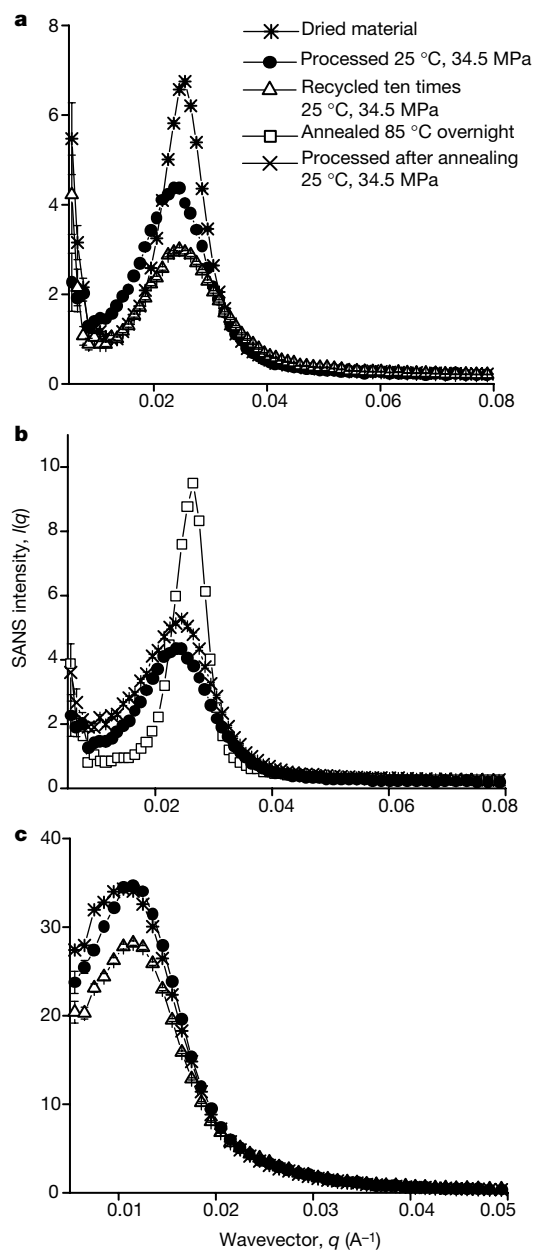


Figure 2 Small-angle neutron scattering intensity, $I(q)$, versus wavevector, q , plots. **a, b**, $60,000 \text{ g mol}^{-1}$ polystyrene-*block*-poly(2-ethylhexyl acrylate), PS-*b*-PEHA, with 52 wt% PEHA and the indicated processing conditions. **c**, PEHA/PS core-shell with 50 wt% PEHA and the indicated processing conditions.

Core-shell nanoparticles with one glassy and one rubbery component have been exploited as coatings^{24,25} and as impact-modifying additives for plastics (with a crosslinked rubbery core)²⁶. To create systems that could be moulded as bulk plastics by pressure-induced mixing, we synthesized core-shell nanoparticles consisting of a non-crosslinked, low- T_g core polymer and a high- T_g shell polymer.

The baroplastic properties of PEHA/PS core-shell nanoparticles with 50 wt% polystyrene and average particle diameter of 67 nm are demonstrated in Fig. 1c, depicting a compression-moulded specimen processed at 34.5 MPa for 5 min at 25 °C from the dried state and one reprocessed ten times following the procedure described above. Similar to the PS-*b*-PEHA block copolymer, the core-shell material can be readily remoulded at room temperature after

repeated processing operations, an effect that we propose derives from the pressure-induced partial miscibilization of the PEHA and PS components.

SANS data for the PEHA/PS nanoparticle system in the as-dried state, after one moulding operation and after ten recycles are shown in Fig. 2c. Several important features are notable. First, each data set exhibits a broad maximum characteristic of the interparticle spacing, suggesting that, similar to block copolymer baroplastics, the initial sample morphology is substantially preserved during low-temperature processing. Upon the initial moulding, the peak position shifts to a slightly larger wavevector. This behaviour was found for all the core-shell systems studied and reflects the material's densification under compression. With further processing, the peak position remains constant but its amplitude diminishes, suggesting enhanced mixing and a corresponding loss of contrast between PS and PEHA domains. This latter result is strong evidence that PS/PEHA indeed exhibits the predicted pressure-induced miscibility.

DSC data on this core-shell system are consistent with a semi-solid processing mechanism. Figure 3c shows DSC traces for the as-dried material and the specimen recycled ten times. Three glass transitions are apparent in each data set, at -65, 93 and 64 °C, corresponding to the PEHA, PS and interphase regions, respectively. (A fourth T_g may be weakly present for both samples near -35 °C.) The fact that little variation is observed in these traces after ten recycles suggests that the system does not undergo wholesale mixing under pressure, but instead flows and forms as a semi-solid.

Although processing kinetics were not a focus of this initial investigation, a preliminary study of the effects of processing time on morphology were performed on a PBA/PS- d_8 core-shell nanoparticle system with 66 wt% PS- d_8 and an average particle diameter of 87 nm. SANS data for this system processed at 34.5 MPa and 25 °C for 1, 10 and 30 min show a reduction in the peak intensity with time (not shown), indicating the PS- d_8 and PBA components become increasingly intermixed. DSC traces for this system processed for 1 and 30 min again show three T_g values near -40, 90 and 47 °C, corresponding to the PBA, PS and interphase regions of the particles, respectively. For this material, the mixed-state T_g can be estimated from the components T_g and their mass fractions, w , from $1/T_{g,mix} = w_1/T_{g,1} + w_2/T_{g,2}$ to be 28 °C — effectively eliminating a processing mechanism that involves complete mixing of the two components.

As controls, PS/PI block copolymers and core-shell nanoparticles were also investigated, to confirm that the low-temperature processing capability witnessed in our investigations had its origins in the pressure-enhanced miscibility of the systems studied. The application of pressure is known to reduce the miscibility of PS/PI^{16–18}, allowing us to separate structural and kinetic effects from thermodynamics. Attempts to compression-mould a commercial PS-*b*-PI-*b*-PS triblock copolymer with 22 wt% PS and $M_n = 90,000 \text{ g mol}^{-1}$ at 25 °C and 55 °C were unsuccessful. Interestingly, a PI/PS core-shell nanoparticle system having 62 wt% PS and total average diameter of 29 nm could be processed into transparent films from the dried precipitate; however, the product was easily fractured and showed poor cohesion during attempts to reprocess.

In summary, we have demonstrated moulding of baroplastic block copolymers and core-shell nanoparticles comprising one glassy and one rubbery component, solely by applying pressure. Capitalizing on pressures typically experienced in plastics manufacturing operations such as injection- and compression-moulding, baroplastics processing may be feasible using current manufacturing equipment. The molecular mechanism underlying this phenomenon appears to involve pressure-induced partial intermixing of dissimilar nanophase domains, resulting in a semi-solid state that facilitates moulding. Such a semi-solid forming operation could be compared to semi-solid metals casting processes that allow for rapid, high-precision alloy castings from slurries or billets

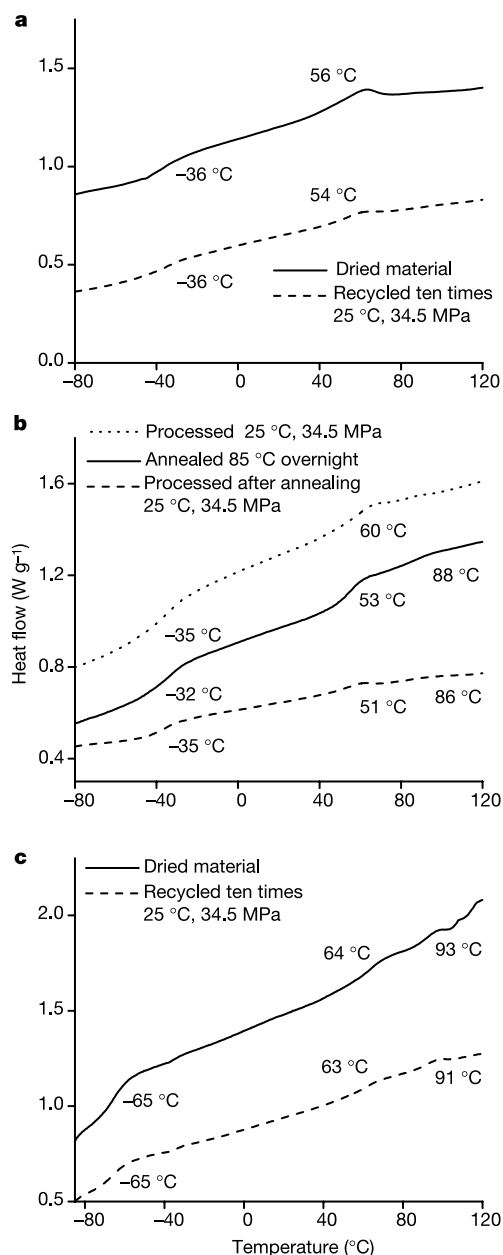


Figure 3 Heat flow versus temperature traces from differential scanning calorimetry, DSC. **a**, **b**, polystyrene-*b*-poly(2-ethylhexyl acrylate), PS-*b*-PEHA, with the indicated processing conditions. **c**, PEHA/PS core-shell with 50 wt% PEHA and the indicated processing conditions. Measurements were performed on a TA Instruments Q1000 DSC at a 5 °C min⁻¹ rate.

with solids content as high as 40–50% (ref. 27).

As a new paradigm for plastics manufacturing, semi-solid processing of baroplastics raises prospects for the low-temperature moulding of plastics incorporating high- T_g , crystallizing or inorganic components, wherein such hard phases move within a fluidized baroplastic medium. We are currently investigating such systems. By removing the resin-heating and mould-cooling requirements, pressure-based processing could decrease both energy consumption and manufacturing time, while eliminating the thermooxidative degradation that limits plastics recyclability. □

Methods

Block copolymer synthesis

Block copolymers of PS-*b*-PBA and PS-*b*-PEHA were synthesized by atom transfer radical polymerization (ATRP)^{20,28,29}. The polystyrene block was first polymerized using methyl-2-bromo-propionate as initiator and CuCl/N,N,N',N',N''-pentamethyldiethylene triamine as the catalyst complex in toluene solution at 100 °C. Once the styrene polymerization reached completion, the temperature was lowered to 80 °C and acrylate monomer was added to obtain the second block. The resulting polymer solution was then passed through an alumina column to remove the catalyst and the copolymer precipitated in methanol. The recovered block copolymers were purified by repeated dissolution in dichloromethane followed by precipitation in methanol. The materials were then dried under vacuum followed by freeze-drying from benzene overnight. Compositions and molecular weights of the resulting block copolymers were determined using ¹H nuclear magnetic resonance (NMR) and gel permeation chromatography (GPC) based on polystyrene standards.

Core-shell nanoparticle synthesis

Core-shell nanoparticles of PEHA/PS and PBA/PS-*d*₈ were synthesized by a two-stage microemulsion polymerization technique^{23,26}. Tetradecyltrimethylammonium bromide and 2,2'-azobis(2-methylpropionamide) dihydrochloride were used as emulsifier and initiator, respectively. Polymerization was performed under nitrogen at 65 °C. Acrylate monomer was first added slowly to ionized water in the presence of emulsifier with vigorous stirring and reacted for 15 h. Pre-emulsified styrene was then added slowly to this solution and allowed to react for 3 h. The resulting core-shell particles were precipitated in methanol/water (with a trace amount of NaCl) and washed in DI water several times. The product was then vacuum-dried in the presence of phosphorus pentoxide for three days at room temperature. Compositions and molecular masses were determined as above. Average particle sizes were determined by dynamic light scattering using a Brookhaven Instruments Co. Zeta Potential Analyzer fitted with a 676-nm laser source.

SANS

Measurements were performed at The Manuel Lujan Jr. Neutron Scattering Center at Los Alamos National Laboratory on the Low- Q diffractometer, LQD, with the following instrument configuration: wavelength = 1.5–15 Å at 20 Hz, scattering angle = 6–60 mrad on a 59-cm diameter detector, resulting in a q range of 0.003–0.5 Å⁻¹. Samples were ~1-cm-diameter disks of variable thickness. Scattered intensities were corrected for background and thickness in the standard manner.

Received 21 March; accepted 21 October 2003; doi:10.1038/nature02140.

- Herbst, H., et al. in *Frontiers in the Science and Technology of Recycling* (ed. Akovali, G.) 73–101 (Kluwer Academic, Dordrecht, The Netherlands, 1997).
- Pollard, M., Russell, T. P., Ruzette, A. V., Mayes, A. M. & Gallot, Y. The effect of hydrostatic pressure on the lower critical ordering transition in diblock copolymers. *Macromolecules* **31**, 6493–6498 (1998).
- Ruzette, A.-V. G., Mayes, A. M., Pollard, M., Russell, T. P. & Hammouda, B. Pressure effects on the phase behavior of styrene/*n*-alkyl methacrylate block copolymers. *Macromolecules* **36**, 3351–3356 (2003).
- Ruzette, A.-V. G., Banerjee, P., Mayes, A. M. & Russell, T. P. A simple model for baroplastic behavior in block copolymer melts. *J. Chem. Phys.* **114**, 8205–8209 (2001).
- Ryu, D. Y., Lee, D. J., Kim, J. K., Lavery, K. A. & Russell, T. P. Effect of hydrostatic pressure on closed-loop phase behavior of block copolymers. *Phys. Rev. Lett.* **90**, 235501 (2003).
- Hasegawa, H. et al. Small-angle neutron scattering studies on phase behavior of block copolymers. *J. Phys. Chem. Solids* **60**, 1307–1312 (1999).
- Frielinghaus, H., Schwahn, D., Mortensen, K., Almdal, K. & Springer, T. Composition fluctuations and coil conformations in a poly(ethylene-propylene)-poly(ethyl ethylene) diblock copolymer as a function of temperature and pressure. *Macromolecules* **29**, 3263–3271 (1996).
- Schwahn, D., Frielinghaus, H., Mortensen, K. & Almdal, K. Temperature and pressure dependence of the order parameter fluctuations, conformational compressibility, and the phase diagram of the PEP-PDMS diblock copolymer. *Phys. Rev. Lett.* **77**, 3153–3156 (1996).
- Russell, T. P., Karis, T. E., Gallot, Y. & Mayes, A. M. Lower critical ordering transition in a diblock copolymer melt. *Nature* **368**, 729–731 (1994).
- Ryu, D. Y., Jeong, U., Kim, J. K. & Russell, T. P. Closed-loop phase behaviour in block copolymers. *Nature Mater.* **1**, 114–117 (2002).
- Holden, G. et al. in *Thermoplastic Elastomers* (ed. Holden, G.) 573–599 (Hanser, Munich/Vienna/New York, 1996).
- Ehrenstein, G. H. *Polymeric Materials* 63–89, 98–116 (Hanser, Munich, 2001).
- Ruzette, A.-V. G. & Mayes, A. M. A simple free energy model for weakly interacting polymer blends. *Macromolecules* **34**, 1894–1907 (2001).
- Gonzalez-Leon, J. A. & Mayes, A. M. Phase behavior prediction of ternary polymer mixtures. *Macromolecules* **36**, 2508–2515 (2003).
- Ruzette, A.-V. G. et al. Phase behavior of diblock copolymers between styrene and *n* alkyl

- methacrylates. *Macromolecules* **31**, 8509–8517 (1998).
- Hajduk, D. A., Urayama, P., Gruner, S. M. & Erramilli, S. High-pressure effects on the disordered phase of block copolymer melts. *Macromolecules* **28**, 7148–7156 (1995).
- Hajduk, D. A., Gruner, S. M., Erramilli, S., Register, R. A. & Fetters, L. J. High-pressure effects on the order/disorder transition in block copolymer melts. *Macromolecules* **29**, 1473–1481 (1996).
- Steinhoff, B. et al. Pressure dependence of the order-to-disorder transition in polystyrene/polyisoprene and polystyrene/poly(methylphenylsiloxane) diblock copolymers. *Macromolecules* **31**, 36–40 (1998).
- Migler, K. B. & Han, C. C. Static and kinetic study of a pressure-induced order-disorder transition: birefringence and neutron scattering. *Macromolecules* **31**, 300–305 (1998).
- Patten, T. E. & Matyjaszewski, K. Atom transfer radical polymerization and the synthesis of polymeric materials. *Adv. Mater.* **10**, 901–915 (1998).
- Malmstrom, E. E. & Hawker, C. J. Macromolecular engineering via 'living' free radical polymerizations. *Macromol. Chem. Phys.* **199**, 923–935 (1998).
- Hillmyer, M. Block copolymer synthesis. *Curr. Opin. Solid State Mater. Sci.* **4**, 559–564 (1999).
- Ha, J. W., Park, I. J., Lee, S. B. & Kim, D. K. Preparation and characterization of core-shell particles containing perfluoroalkyl acrylate in the shell. *Macromolecules* **35**, 6811–6818 (2002).
- Keddie, J. L. Film formation of latex. *Mater. Sci. Eng. R* **21**, 101–170 (1997).
- Dos Santos, F. D. & Leibler, L. Large deformation films from soft-core/hard-shell hydrophobic latexes. *J. Polym. Sci. B* **41**, 224–234 (2003).
- Lovell, P. A. & Pierre, D. in *Emulsion Polymerization and Emulsion Polymers* (eds Lovell, P. A. & Aasser, M. S.) 657–695 (John Wiley & Sons, New York, 1997).
- Flemings, M. C. Behavior of metal alloys in the semisolid state. *Metall. Trans. B* **22**, 269–293 (1991).
- Shipp, D. A., Wang, J.-L. & Matyjaszewski, K. Synthesis of acrylate and methacrylate block copolymers using atom transfer radical polymerization. *Macromolecules* **31**, 8005–8008 (1998).
- Cassebras, M., Pascual, S., Polton, A., Tardi, M. & Vairon, J. P. Synthesis of di- and triblock copolymers of styrene and butyl acrylate by controlled atom transfer radical polymerization. *Macromol. Rapid Commun.* **20**, 261–264 (1999).

Acknowledgements We acknowledge the support of the Seaver Institute, the Lord Foundation, Lord Corporation, the MRSEC Program of the National Science Foundation and the Office of Naval Research. This work benefited from the use of the Los Alamos Neutron Science Center at the Los Alamos National Laboratory, funded by the US Department of Energy.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.M.M. (amayes@mit.edu).

Synthetic design of crystalline inorganic chalcogenides exhibiting fast-ion conductivity

Nanfeng Zheng¹, Xianhui Bu^{2*} & Pingyun Feng¹

¹Department of Chemistry, University of California, Riverside, California 92521, USA

²Department of Chemistry, University of California, Santa Barbara, California 93106, USA

* Present address: Department of Chemistry and Biochemistry, California State University, 1250 Bellflower Blvd, Long Beach, California 90840, USA

Natural porous solids such as zeolites are invariably formed with inorganic cations such as Na⁺ and K⁺ (refs 1, 2). However, current research on new porous materials is mainly focused on the use of organic species as either structure-directing or structure-building units; purely inorganic systems have received relatively little attention in exploratory synthetic work^{3–9}. Here we report the synthesis of a series of three-dimensional sulphides and selenides containing highly mobile alkali metal cations as charge-balancing extra-framework cations. Such crystalline inorganic chalcogenides integrate zeolite-like architecture with high anionic framework polarizability and high concentrations of mobile cations. Such structural features are particularly desirable for the development of fast-ion conductors¹⁰. These materials demonstrate high ionic conductivity (up to $1.8 \times 10^{-2} \text{ ohm}^{-1} \text{ cm}^{-1}$) at room temperature and moderate to high humidity. This synthetic methodology, together with novel

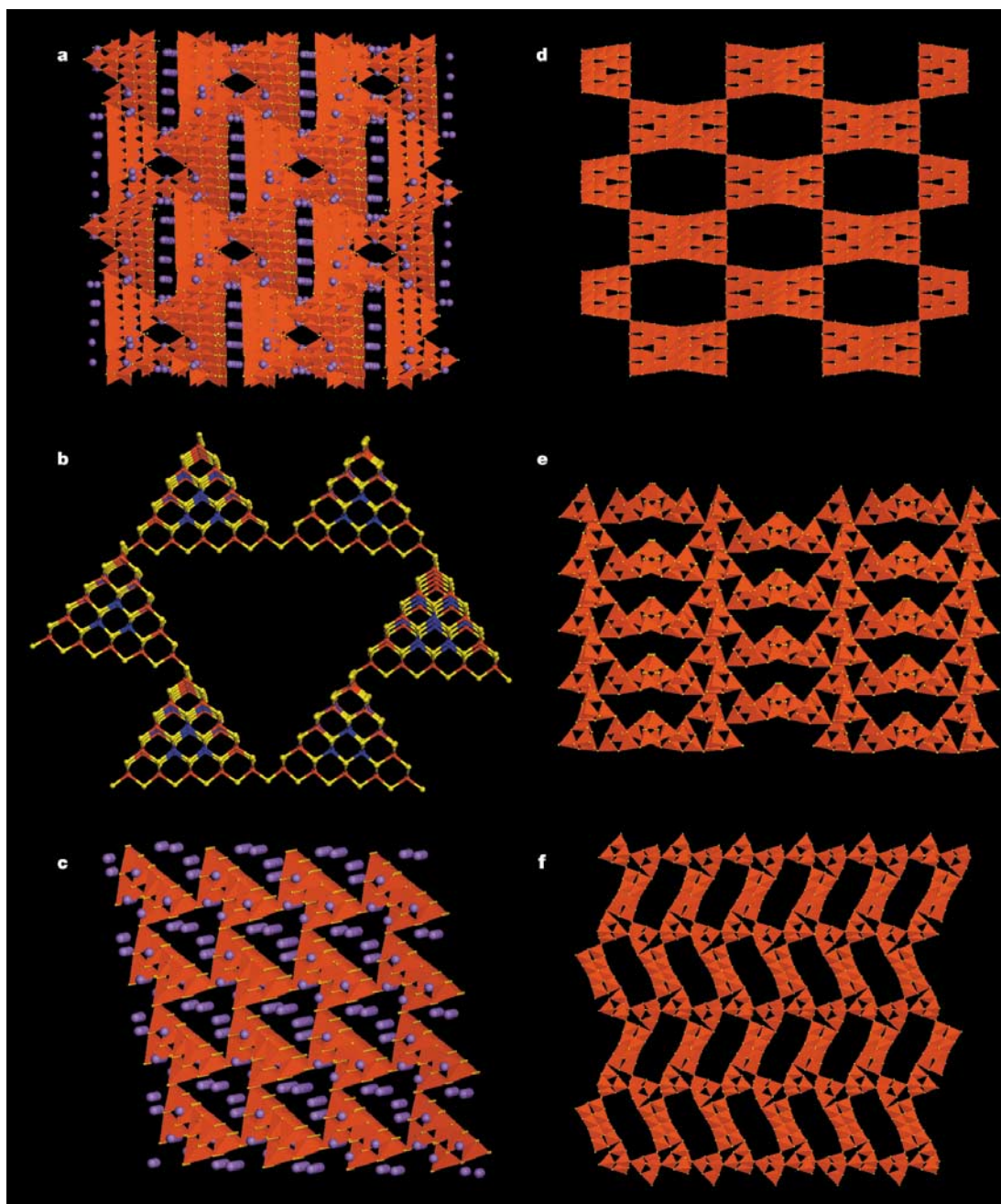


Figure 1 The structural diagrams of the inorganic chalcogenide frameworks. **a**, The 3D diamond-type superlattice in ICF-5 CuInS-Na. Unconnected spheres are Na^+ sites. **b**, The rings formed by six T5 clusters are present in the 3D interpenetrating diamond-type superlattice in ICF-17. **c**, T2 clusters in ICF-21 InSe-Na are connected into a non-

interpenetrating diamond-type lattice. Unconnected spheres are Na^+ sites. **d**, The 3D diamond-type superlattice with core-less T4 clusters ($\text{In}_{16}\text{S}_{32}^{16-}$) in ICF-22 InS-Li. Only one sublattice is shown here for clarity. **e**, The 3D framework of ICF-24 with 20-ring channels. **f**, The 3D framework of ICF-25 with 16-ring channels.

structural, physical and chemical properties, may lead to the development of new microporous and open-framework materials with potential applications in areas such as batteries, fuel cells, electrochemical sensors and photocatalysis.

Commercial applications continue to be dominated by porous materials (for example, zeolites A and X) made from purely inorganic systems. New synthetic methodologies that allow the creation of porous materials from purely inorganic systems will hold great promise for technological applications. One property that could benefit from the compositional and topological features of purely inorganic open-framework chalcogenides is fast-ion conductivity at low temperatures ($<100^\circ\text{C}$). Fast-ion conductors are useful as electrode materials or solid electrolytes in electrochem-

ical devices such as batteries, fuel cells and sensors. Solids generally have negligible ionic conductivity. Most fast-ion conductors discovered so far have a high conductivity only at high temperatures ($>100^\circ\text{C}$)¹⁰. Therefore, the creation of new fast-ion conductors that operate at low temperatures is desirable.

An open-framework material has an inherent advantage for applications such as low-temperature fast-ion conductors because its open channels provide paths for easy ion migration. Unfortunately, zeolites, although they have open channels and cages, are not good fast-ion conductors because of the strong interaction between their oxygen framework and extra-framework charge carriers, such as Li^+ and Na^+ (ref. 2). Open-framework chalcogenides are anticipated to be better ion conductors than zeolites because chalcogen-

Table 1 A summary of crystallographic data for selected synthesized chalcogenides

Name	Framework composition	Space group	<i>a</i> (Å)	<i>b</i> (Å)	<i>c</i> (Å)	<i>R</i> (<i>F</i>)*
ICF-5 CuInS-Na	Cu ₃ In ₁₇ S ₃₃ ¹⁰⁻	<i>Fddd</i>	31.329(9)	32.906(9)	44.546(12)	6.45
ICF-5 CdInS-Na	Cd ₄ In ₁₆ S ₃₃ ¹⁰⁻	<i>I4₁/acd</i>	26.602(4)	26.602(4)	42.660(9)	6.94
ICF-5 MnZnInS-Na	Mn _{11.8} Zn _{2.2} In ₁₆ S ₃₃ ¹⁰⁻	<i>I4₁/acd</i>	26.288(8)	26.288(8)	42.637(18)	8.37
ICF-5 CdInS-Li	Cd ₄ In ₁₆ S ₃₃ ¹⁰⁻	<i>I4₁/acd</i>	25.458(2)	25.458(2)	43.471(4)	12.6
ICF-5 ZnInS-Na	Zn ₄ In ₁₆ S ₃₃ ¹⁰⁻	<i>I4₁/acd</i>	26.158(10)	26.158(10)	42.63(2)	13.9
ICF-5 MnInS-Li	Mn ₄ In ₁₆ S ₃₃ ¹⁰⁻	<i>I4₁/acd</i>	24.690(2)	24.690(2)	44.488(5)	7.20
ICF-17 InZnS-Na	In ₂₂ Zn ₁₃ S ₅₄ ¹⁶⁻	<i>I4₁/acd</i>	29.004(9)	29.004(9)	54.53(2)	9.39
ICF-21 InSe-Na	In ₄ Se ₈ ⁴⁻	<i>I-42d</i>	11.765(2)	11.765(2)	22.685(4)	5.55
ICF-22 InS-Li	In ₄ S ₈ ⁴⁻	<i>I4₁/acd</i>	25.034(7)	25.034(7)	43.506(15)	7.06
ICF-24 InSe-Na	In ₄ S _{2.9} Se _{5.1} ⁴⁻	<i>Fdd2</i>	28.842(7)	50.035(11)	12.141(3)	6.64
ICF-25 InS-SrCaLi	In ₄ S ₈ ⁴⁻	<i>Fdd2</i>	28.33(1)	70.31(2)	12.303(4)	6.97
ICF-27 InS-SrLi	In ₁₅ S ₂₉ ¹³⁻	<i>P3₁c</i>	17.846(3)	17.846(3)	17.333(4)	4.06

**R*(*F*) = Σ||*F*_o| - |*F*_c||/Σ|*F*_o| with *F*_o > 4.0σ(*F*_o).

ides have higher anionic framework polarizability, owing to the large size of S²⁻ or Se²⁻ as compared to O²⁻. A more polarizable anionic framework will facilitate the migration of mobile cations¹⁰. Another advantage of open-framework chalcogenides is the high concentration of mobile cations. Chalcogenides can have much more negative frameworks and therefore more charge-balancing cations than zeolites because the framework M⁴⁺/M³⁺ (where M is a tetrahedral atom) ratio in chalcogenides can be much smaller than one, whereas it is always larger or equal to one in zeolites or related oxides¹¹.

Unfortunately, purely inorganic chalcogenides with the zeolite-like architecture do not occur naturally. The limitation of known open-framework chalcogenides is that they are all made by using organic structure-directing agents^{12–14}. We have now developed a kinetically controlled synthesis method that has led to the synthesis of hydrated sulphides and selenides with highly mobile alkali or alkaline earth metal cations (for example, Li⁺, Na⁺, Ca²⁺) as extra-framework cations (Table 1). Their unprecedented compositions, diverse cluster and crystal structures, and fast-ion conductivity are reported below.

All phases were synthesized from organic-free aqueous solutions under hydrothermal conditions at temperatures below 200 °C (see Methods). As for the synthesis of zeolites, reactions were carried out under highly alkaline conditions. The kinetically controlled processes led to the crystallization of chalcogenides encapsulating mobile inorganic cations in their cavities. These chalcogenides are denoted ICF-*m* (ICF, inorganic chalcogenide framework; *m*, a number related to the type of framework topology). The *m* is



Figure 2 The 3D framework of ICF-27 built from octahedral cluster units.

often followed by elemental symbols (for example, ICF-21 InSe-Na), indicating in which of several possible framework and extra-framework compositions a particular topology has been made.

Compared to zeolite tetrahedral frameworks, these chalcogenides represent a higher level of the structural hierarchy because they can be derived from simple tetrahedral networks (for example, diamond) by replacing tetrahedral sites with supertetrahedral clusters. Supertetrahedral clusters are regular tetrahedrally shaped fragments of the cubic ZnS-type lattice and are denoted as T*n*, where *n* refers to the number of metal layers^{15–17}.

A number of 3D framework types have been realized. Here we focus on seven three-dimensional (3D) framework types, denoted ICF-5, ICF-17, ICF-21, ICF-22, ICF-24, ICF-25 and ICF-27. Because of the compositional variations among each topology, over a dozen

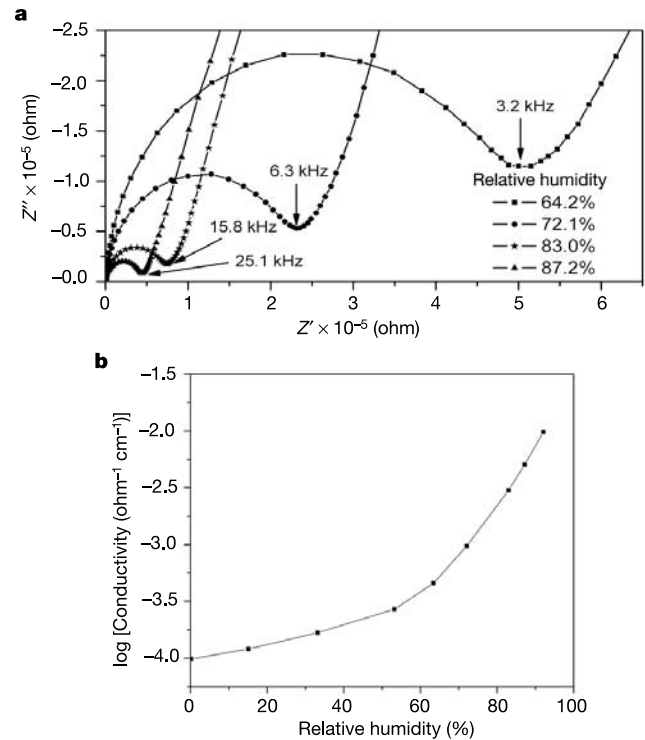


Figure 3 The dependence of the ionic conductivity on the relative humidity for ICF-5 CuInS-Na. **a**, Complex impedance plots under different relative humidities. The data were measured on a single crystal sample at 18 °C. Before each measurement, the sample was kept under the given humidity for 2 h. *Z'* and *Z''* are the real and imaginary parts of the impedance, respectively. **b**, The ionic conductivity of the same sample under different relative humidities.

different materials have been prepared and structurally determined by single-crystal X-ray diffraction (Supplementary Information). That the large supertetrahedral T4 (in ICF-5, Fig. 1a) and T5 (in ICF-17, Fig. 1b) clusters could be prepared from inorganic systems is an important step towards the development of novel clusters. Before this work, chalcogenide clusters were usually synthesized in the presence of organic ligands or bases that serve either to stabilize the cluster surface or to balance the charge on clusters.

In ICF-5, T4 clusters ($M_{20}S_{33}^{10-}$) are joined together into two interpenetrating lattices (Fig. 1a). ICF-5 can be made in diverse chemical compositions with either lithium or sodium as extra-framework cations (Table 1). ICF-17 contains the largest supertetrahedral cluster (for example, $Zn_{13}In_{22}S_{54}^{16-}$) made from M^{2+} and M^{3+} cations (Fig. 1b). In this cluster, M^{2+} cations are distributed at the core and faces of the cluster whereas higher-charged In^{3+} cations terminate the cluster surface at the edges and corners of the cluster. These T5 clusters resemble the core-shell nanoparticles. ICF-21 InSe-Na (formula $NaInSe_2 \cdot xH_2O$) with the T2 cluster $In_4Se_8^{4-}$ is the first hydrated selenide with the 4-connected, 3D framework (Fig. 1c). Its structure is distinct from the anhydrous $NaInSe_2$ that has a layered structure with octahedral In^{3+} cations¹⁸.

In ICF-22, core-less T4 clusters ($In_{16}S_{32}^{16-}$) are assembled into two interpenetrating lattices (Fig. 1d). The core-less T4 cluster is similar to the ordinary T4 cluster (for example, $Zn_4In_{16}S_{33}^{10-}$) except that the core sulphur atom and its adjacent four metal (for example, Zn_4S^{6+}) atoms are missing.

One prominent structural feature in ICF-24 and ICF-25 is the presence of rings with three T2 clusters. At the other extreme, rings of ten T2 clusters in ICF-24 represent the largest-known ring size

formed by supertetrahedral clusters (Fig. 1e). ICF-24 is the first 4-connected, 3D framework structures with a ring size of 20. In comparison, the ring size in ICF-25 is 16, made of eight T2 clusters (Fig. 1f). As in ICF-21, both ICF-24 and ICF-25 have 3D non-centrosymmetric framework structures. ICF-27 can be considered as a 6-connected network with the $In_{18}S_{41}^{28-}$ cluster as the pseudo-octahedral building unit. The $In_{18}S_{41}^{28-}$ clusters are cross-linked in three dimensions by edge-sharing InS_4 tetrahedra at six corners (Fig. 2).

One notable difference from zeolites is the chemical compositions in which the above topological features are realized. ICF-21 InSe-Na, ICF-22 InS-Li, ICF-24, and ICF-25 are the first examples of zeolite-like, 4-connected, 3D open-framework chalcogenides built without tetravalent cations. They represent extreme cases that demonstrate the fundamental difference between oxides and chalcogenides. Zeolite-like oxides generally follow the Loewenstein rule that states that the ratio of M^{4+}/M^{3+} must be larger than or equal to one.

The non-compliance with the Loewenstein rule can lead to a higher concentration of mobile charge carriers in chalcogenides than in zeolites. For a zeolite, not only are M^{4+} cations required in the framework, but the ratio of M^{4+} to M^{3+} cations has to be larger than or equal to one (for example, $NaAlSiO_4 \cdot xH_2O$). This limits the maximum number of monovalent charge-balancing cations to 0.5 per framework tetrahedral cation. On the other hand, the M^{4+}/M^{3+} ratio in 4-connected, 3D chalcogenides can be less than one (ref. 11). Even the extreme limit ($M^{4+}/M^{3+} = 0$) can be realized, as shown here. The lower M^{4+}/M^{3+} ratio leads to a more negative framework and a higher concentration of extra-framework mobile cations. There can be up to one monovalent charge carrier per tetrahedral cation in chalcogenides (compared to the maximum value of 0.5 in zeolites and related oxides). A high concentration of charge carriers will help enhance ionic conductivity.

The integration of the open-framework architecture with high framework polarizability and high concentration of charge carriers gives rise to a high ionic conductivity in these chalcogenides. The specific conductivity of ICF-5 CuInS-Na is $1.2 \times 10^{-2} \text{ ohm}^{-1} \text{ cm}^{-1}$ at 18 °C under 100% relative humidity (Fig. 3a, b) (see Methods). A high value is also observed in the selenide ICF-21 InSe-Na with the specific conductivity of $3.4 \times 10^{-2} \text{ ohm}^{-1} \text{ cm}^{-1}$ at 21 °C under 100% relative humidity (Fig. 4a, b). Of particular interest is the high ionic conductivity observed in one lithium-containing

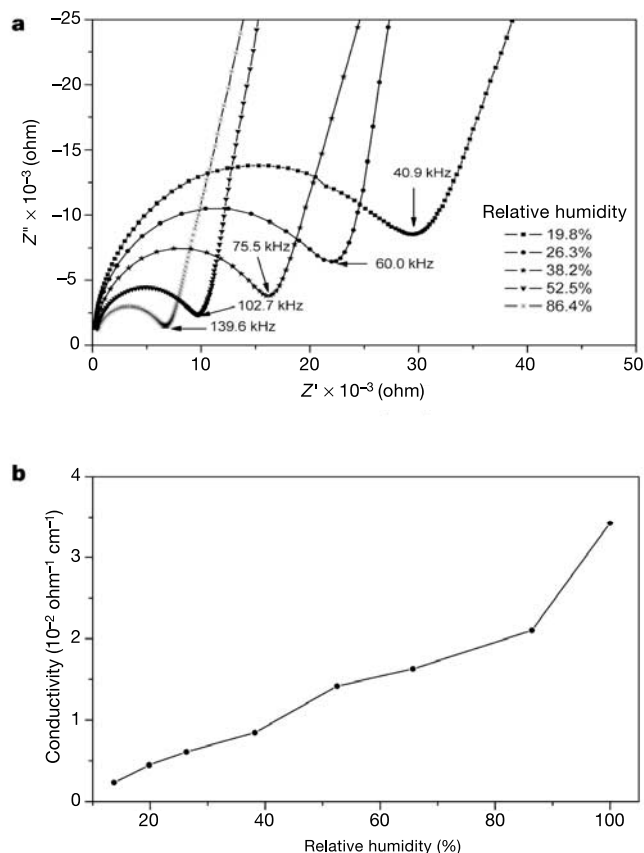


Figure 4 The dependence of the ionic conductivity on the relative humidity for ICF-21 InSe-Na. **a**, Complex impedance plots at 21 °C under different relative humidities. **b**, The ionic conductivity under different relative humidities.

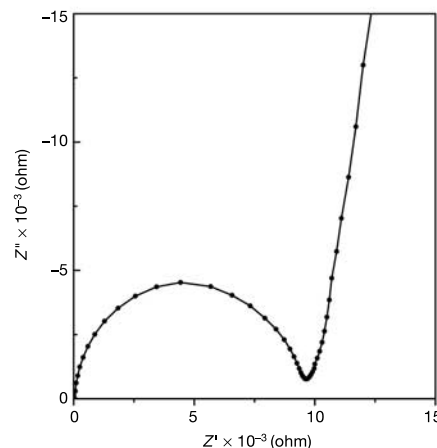


Figure 5 The a.c. impedance plot of ICF-22 InS-Li at 24.0 °C under 31.7% relative humidity. The area of the cross-section for the pellet is $3.83 \times 10^{-3} \text{ cm}^2$ and the length is 0.405 cm. The sample resistance from this plot is 9572.2 Ω. The specific conductivity is $1.1 \times 10^{-2} \text{ ohm}^{-1} \text{ cm}^{-1}$.

material, ICF-22 InS-Li (Supplementary Information). Its specific conductivity reaches up to $1.1 \times 10^{-2} \text{ ohm}^{-1} \text{ cm}^{-1}$ at 24 °C under 31.7% relative humidity (Fig. 5). The ionic conductivity of these chalcogenides generally increases with the relative humidity. For example, at 18 °C, the conductivity of ICF-5 CuInS-Na ranges from $1.0 \times 10^{-4} \text{ ohm}^{-1} \text{ cm}^{-1}$ under 0.2% relative humidity to $1.2 \times 10^{-2} \text{ ohm}^{-1} \text{ cm}^{-1}$ under 100% relative humidity (Fig. 3a, b). A similar trend is observed for ICF-21 InSe-Na (Fig. 4a, b).

Despite their high ionic conductivity, there are some limitations for the practical applications of these materials, particularly for lithium batteries, because a relative humidity of 30% or higher is needed to achieve the highest ionic conductivity. Another possible application might take advantage of the open architecture and narrow bandgaps (as compared to oxides) of these materials for photocatalysis. In either case, further research is needed to develop these materials for practical applications.

The materials reported are probably only a small group of a much larger family of inorganic open-framework chalcogenides. The synthetic method described here should be applicable to a range of chalcogenide structures and compositions. The diverse chemical compositions and crystal structures provide rich opportunities in exploring the physical and chemical properties of these chalcogenides and in establishing structure-property relationships. □

Methods

Typical synthesis conditions are given below using ICF-22 InS-Li, ICF-21 InSe-Na, and ICF-5 CuInS-Na as examples.

For ICF-22 InS-Li, $\text{In}(\text{NO}_3)_3 \cdot \text{H}_2\text{O}$ (339.2 mg), LiCl (176.9 mg), Li_2S (210.3 mg), and H_2O (2.1967 g) were mixed and stirred in a 23-ml Teflon-lined stainless steel autoclave for 20 min. The vessel was then sealed and heated at 190 °C for 4 days. After cooling to room temperature, a ~68% yield of colourless crystals was obtained.

For ICF-21 InSe-Na, a mixture of $\text{In}(\text{NO}_3)_3 \cdot \text{H}_2\text{O}$ (381.5 mg), Na₂Se (408.3 mg), and H_2O (2.0240 g) was prepared and stirred in a 23-ml Teflon-lined stainless steel autoclave for 10 min. The vessel was then sealed and heated at 170 °C for 3 days. The autoclave was cooled to room temperature. A ~77% yield of pale-yellow octahedral crystals was obtained.

For ICF-5 CuInS-Na, $\text{In}(\text{NO}_3)_3 \cdot \text{H}_2\text{O}$ (372.0 mg), $\text{Cu}(\text{NO}_3)_2 \cdot 3\text{H}_2\text{O}$ (67.98 mg), Na₂S (234.4 mg), and water (2.5740 g) were mixed in a 23-ml Teflon-lined stainless steel autoclave and stirred for 10 min. The vessel was then sealed and heated at 150 °C for 3 days. After cooling to room temperature, a ~85% yield of red crystals was obtained.

For ICF-5 CuInS-Na, the elemental analysis (wt%) was 44.22 In, 4.03 Cu (calculated 43.97 In, 4.29 Cu) based on the formula $\text{Na}_{10}[\text{Cu}_3\text{In}_7\text{S}_{33}] \cdot 56(\text{H}_2\text{O})$. The thermal analysis shows that ICF-5 CuInS-Na undergoes a total of 17.0% water loss in the temperature range from room temperature to 200 °C. All crystal structures were solved from single crystal data collected at 150 K on a SMART 1000 charge-coupled device (CCD) diffractometer.

Ionic conductivities were measured by a.c. impedance methods with an applied frequency range of 10 to 20 MHz using a Solartron 1260 frequency response analyser. Both pellet and single crystal samples were used for measurements. Pellets were prepared by compressing the fresh powder sample at the pressure of 5,000 psi and were then cut into smaller pellets shaped like the rectangular prism. Liquid gallium was used as contacting electrodes.

Received 4 May; accepted 27 October 2003; doi:10.1038/nature02159.

- Flanigen, E. M. in *Introduction to Zeolite Science and Practice* (eds van Bekkum, H., Flanigen, E. M. & Jansen, J. C.) 13–34 (Elsevier, New York, 1991).
- Breck, D. W. *Zeolite Molecular Sieves, Structure, Chemistry, and Use* (John Wiley & Sons, New York, 1974).
- Davis, M. E. Ordered porous materials for emerging applications. *Nature* **417**, 813 (2002).
- Feng, P., Bu, X. & Stucky, G. D. Hydrothermal syntheses and structural characterization of zeolite analogue compounds based on cobalt phosphate. *Nature* **388**, 735–741 (1997).
- Scott, R. W. J., MacLachlan, M. J. & Ozin, G. A. Synthesis of metal sulfide materials with controlled architecture. *Curr. Opin. Solid State Mater. Sci.* **4**, 113–121 (1999).
- Huo, Q., Leon, R., Petroff, P. M. & Stucky, G. D. Mesoporous design with gemini surfactants: supercage formation in a three-dimensional hexagonal array. *Science* **268**, 1324–1327 (1995).
- Zhao, D. et al. Triblock copolymer syntheses of mesoporous silica with periodic 50 to 300 angstrom pores. *Science* **279**, 548–552 (1998).
- Johnson, S. A., Ollivier, P. J. & Mallouk, T. E. Ordered mesoporous polymers of tunable pore size from colloidal silica templates. *Science* **283**, 963–965 (1999).
- Cheetham, A. K., Ferey, G. & Loiseau, T. Open-framework inorganic materials. *Angew. Chem. Int. Edn* **38**, 3268–3292 (1999).
- West, A. R. *Solid State Chemistry and its Applications* (Wiley, New York, 1992).
- Zheng, N., Bu, X. & Feng, P. Microporous and photoluminescent chalcogenide zeolite analogs. *Science* **298**, 2366–2369 (2002).
- Bedard, R. L., Wilson, S. T., Vail, L. D., Bennett, J. M. & Flanigen, E. M. in *Zeolites: Facts, Figures, Future. Proc. 8th Int. Zeolite Conf.* (eds Jacobs, P. A. & van Santen, R. A.) 375 (Elsevier, Amsterdam, 1989).
- Cahill, C. L. & Parise, J. B. On the formation of framework indium sulfides. *J. Chem. Soc. Dalton Trans.* 1475–1482 (2000).

- Dhingra, S. & Kanatzidis, M. G. Open framework structures based on Se_2^{2-} fragments: synthesis of $(\text{Ph}_4\text{P})[\text{M}(\text{Se}_2)_2]$ (M = Ga, In, Tl) in molten $(\text{Ph}_4\text{P})_2\text{Se}_6$. *Science* **258**, 1769–1772 (1992).
- Li, H., Laine, A., O’Keeffe, M. & Yaghi, O. M. Supertetrahedral sulfide crystals with giant cavities and channels. *Science* **283**, 1145–1147 (1999).
- Wehmschulte, R. J. & Power, P. P. Low-temperature synthesis of aluminum sulfide as the solvate $\text{Al}_4\text{S}_6(\text{NMe}_3)_4$ in hydrocarbon solution. *J. Am. Chem. Soc.* **119**, 9566–9567 (1997).
- Bu, X., Zheng, N., Li, Y. & Feng, P. Pushing up the size limit of chalcogenide supertetrahedral clusters: two- and three-dimensional photoluminescent open frameworks from $(\text{Cu}_3\text{In}_{30}\text{S}_{54})^{13-}$ clusters. *J. Am. Chem. Soc.* **124**, 12646–12647 (2002).
- Hoppe, R., Lidecke, W. & Frotath, F. C. Sodium thioindate and sodium selenoindate. *Z. Anorg. Allgem. Chem.* **309**, 49–54 (1961).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We acknowledge the support of this work by the NSF. We also thank Y. Yan and his group for assistance with impedance measurements.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P. F. (pingyun.feng@ucr.edu).

Explosive volcanism may not be an inevitable consequence of magma fragmentation

Helge M. Gonnermann & Michael Manga

Earth and Planetary Science, University of California, Berkeley, California 94720, USA

The fragmentation of magma, containing abundant gas bubbles, is thought to be the defining characteristic of explosive eruptions^{1–3}. When viscous stresses associated with the growth of bubbles and the flow of the ascending magma exceed the strength of the melt^{2,4–6}, the magma breaks into disconnected fragments suspended within an expanding gas phase. Although repeated effusive and explosive eruptions for individual volcanoes are common^{7,8}, the dynamics governing the transition between explosive and effusive eruptions remain unclear. Magmas for both types of eruptions originate from sources with similar volatile content, yet effusive lavas erupt considerably more degassed than their explosive counterparts^{7,8}. One mechanism for degassing during magma ascent, consistent with observations, is the generation of intermittent permeable fracture networks generated by non-explosive fragmentation near the conduit walls^{9–11}. Here we show that such fragmentation can occur by viscous shear in both effusive and explosive eruptions. Moreover, we suggest that such fragmentation may be important for magma degassing and the inhibition of explosive behaviour. This implies that, contrary to conventional views, explosive volcanism is not an inevitable consequence of magma fragmentation.

Analysis of pumice samples from the explosive eruption of Mount Pinatubo, 15 June 1991, suggests that non-newtonian rheology and growth of gas bubbles during magma ascent led to intense shear and possibly fragmentation at the conduit walls⁹. Recent observations suggest that fragmentation at the conduit walls may not be restricted to explosive eruptions. Tuffen *et al.*¹⁰ describe multiple generations of tuffsite veins in the dissected vent of a rhyolite lava flow at Torfajökull, Iceland. They¹⁰ interpret these features to record repeated cycles of brittle fragmentation followed by annealing of the fragments and subsequent deformation of the reannealed magma by viscous flow, to form flow banding. Textures within the tuffsite veins suggest that fragmentation created a

transient fracture system amenable to gas flow. Similar examples of apparent fragmentation, reannealing and deformation of fragments into flow banding are also observed in the Mule Creek vent, New Mexico¹¹. Non-explosive fragmentation of ascending magma has also been suggested for the generation of volcanic earthquakes

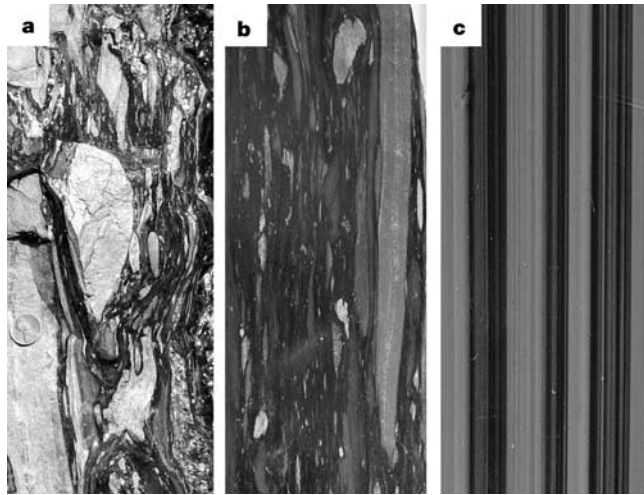


Figure 1 Deformation textures of annealed fragments in obsidian from Big Glass Mountain, California. **a**, Annealed fragments are slightly elongate parallel to shear direction; sample width is 11.5 cm. **b**, Annealed and deformed fragments; sample width is 6 cm. **c**, Typical 'flow banding'; sample width is 4 cm. Fragment compositions in all three samples are identical: the difference in colour is primarily due to varying abundance of micrometre-size crystals.

during the 1991–1995 effusive eruption of the Unzen volcano, Japan¹². We have examined obsidians from the Big Glass Mountain rhyolite dome, California, and find corroborating evidence of magma fragmentation, reannealing and elongation of fragments into flow banding (Fig. 1).

To gain insight into the mechanics governing fragmentation in silicic volcanoes, we have developed a numerical model for magma ascent in the volcanic conduit. The ascending magma (that is, melt + bubbles) is modelled as the steady, isothermal flow of a single-phase liquid at constant mass flux, Q_m , in a cylindrical conduit of constant radius R . We specify a magma pressure, $p_m = p_0$, number density of bubbles, n_d , and relaxed newtonian melt viscosity, μ_{mr} , at the base of the conduit to solve the joint problem of bubble growth and magma ascent. Rather than include the transition to fragmentation and flow of fragmented magma, we determine the ascent distance above the conduit entry at which magma fragmentation by viscous shear should first occur. Unlike previous studies^{3,13}, the model is quasi-one-dimensional and for a given depth computes the radially varying vertical component of magma velocity, $u(r, z)$. Mass and momentum equations solved by our model are:

$$\frac{d}{dz} \left[\rho(z) \int_A u(r, z) dA \right] = \frac{d}{dz} [\rho(z) Q(z)] = \frac{d}{dz} Q_m = 0 \quad (1)$$

and

$$\mu(\dot{\gamma}, \phi, c, a, r, z) \dot{\gamma}(r, z) = -\frac{r}{2} \frac{dp(z)}{dz} \quad (2)$$

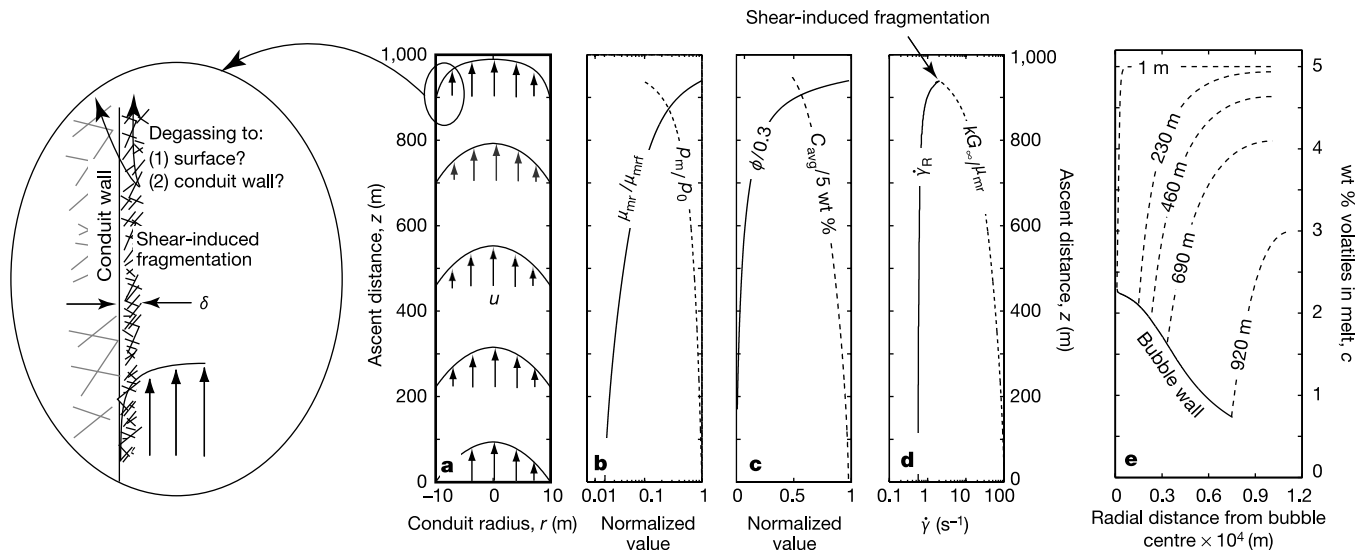


Figure 2 Evolution of a typical model simulation. The model simulation starts at an ambient pressure of 250 MPa and ends once shear-induced fragmentation is first predicted to occur. **a**, Calculated velocity profiles, $u(r)$, of the ascending magma, shown at different ascent distances. The inset is a schematic depiction of the theoretical model for shear-induced magma fragmentation near the conduit walls and concurrent open-system degassing. Small hatch marks schematically indicate shear-induced fragmentation, which for this model calculation is inferred to occur at 920 m. The presence of shear-induced fragmentation above the initial fragmentation level is hypothetical and is not based on actual model calculations. The two large arrows indicate possible degassing pathways associated with the increased permeability of fragmented magma. **b**, Relaxed melt viscosity normalized by its value at the first occurrence of shear-induced fragmentation (μ_{mr}/μ_{mr1} , solid line) and ambient pressure normalized by its value at zero ascent distance (p_m/p_0 , dashed line) versus ascent distance. The gradient in p_m is the

sum of magma-static and dynamic pressure changes. **c**, Normalized magma vesicularity (solid line) and normalized volumetrically averaged dissolved volatile content (dashed line) versus ascent distance. **d**, Calculated radial strain rate at the conduit wall ($\dot{\gamma}_R$, solid line) and critical strain rate for fragmentation (KG_∞/μ_{mr} , dashed line) versus ascent distance. The calculated values of $\dot{\gamma}_R$ increase with ascent distance as the velocity profile becomes increasingly blunt and the largest strain rates are near the conduit walls. The critical strain rate decreases as the relaxed melt viscosity increases with volatile exsolution. When calculated $\dot{\gamma}_R$ and critical strain rate are equal, the first occurrence of shear-induced fragmentation is inferred and the model calculation ends. **e**, Dissolved volatile content of the melt. Each concentration profile (dashed line) corresponds to one velocity profile shown in **a**. Also shown is the change in bubble radius and volatile content at the bubble wall (solid line). Diffusion of volatiles into the gas bubbles, as a consequence of decreasing p_m , results in decreasing volatile content of the melt and increasing μ_{mr} shown in **b**.

with boundary conditions:

$$u(R, z) = \frac{du(0, z)}{dr} = 0 \quad (3)$$

Here A is the cross-sectional area of the conduit, $\rho(z)$ is the magma density, $Q(z) = Q_m/\rho(z)$ is the volumetric flow rate, $\mu(r, z)$ is the magma viscosity, $p(z)$ is the dynamic pressure, r and z are the radial and vertical coordinates respectively, $\dot{\gamma} = du/dr$ is the rate of simple shear strain, $\phi(z)$ is magma vesicularity, $a(z)$ is bubble radius, and $c(z)$ is the concentration of volatiles dissolved in the melt. Calculations of volatile exsolution and bubble growth, which occur as a consequence of depressurization during magma ascent, assume a homogeneous suspension of spherical bubbles surrounded by uniform shells of silicate melt¹⁴ with a volatile-dependent viscosity¹⁵.

During the flow calculations, magma viscosity is allowed to vary vertically and radially. Locally, magma viscosity depends on melt viscosity, μ_m , and strain-rate dependent suspension (melt + bubbles) rheology^{16–18}. Melt viscosity, in turn, depends on strain rate^{4–6,19}, as well as volatile content²⁰. We do not explicitly model the growth of crystals during magma ascent. However, the effect of crystallinity on magma rheology can be represented by increased magma viscosity²¹, in accordance with rheological models for suspensions of rigid particles^{16–18}.

Figure 2 depicts the typical changes in flow conditions and radial velocity profiles (Fig. 2a) calculated by our model. During ascent, magma pressure decreases from its initial value through magmatic and dynamic pressure loss (Fig. 2b) as:

$$p_m = p_0 - \int_0^z \left[\frac{dp}{dz} + g\rho(z) \right] dz \quad (4)$$

This results in volatile exsolution and bubble growth^{14,15} (Fig. 2e), which in turn causes ϕ , Q , and μ_{mr} to increase (Fig. 2b,c). Because of the rheological dependence on $\dot{\gamma}$, there is an inherent feedback between increasing $\dot{\gamma}$ and decreasing μ . The no-slip boundary condition results in large shear-strain rates at the conduit walls, $\dot{\gamma} = \dot{\gamma}_R$, and the initially near-newtonian velocity profile of the ascending magma becomes increasingly blunt¹⁸ (Fig. 2a).

The melt closest to the conduit walls will eventually be subjected to shear-strain rates that exceed the experimentally determined critical strain rate of silicate melts^{2,4,12} (Fig. 2d). In our simulations, the Capillary number, $Ca = \mu_m \dot{\gamma}_R a / \sigma$, is $\gg 1$ near the conduit walls, and we can assume that strain rates of magma and melt are similar²¹. Here σ is surface tension. We adopt a criterion for magma fragmentation:

$$\dot{\gamma} \geq \kappa \frac{G_\infty}{\mu_{mr}} \quad (5)$$

where $\kappa = 0.01$ is an experimental constant and $G_\infty = 10$ GPa is the elastic modulus at infinite frequency.

At the threshold for ‘shear-induced fragmentation’, the melt surrounding gas bubbles is assumed to break by brittle failure^{2–6}. Therefore, we expect that fragmentation depends primarily on the rheology of the melt phase, rather than the magma. We adopt a critical shear-strain rate that is based on the relaxed newtonian melt viscosity, a measured quantity during experimental determination of critical strain rates^{4–6}. However, given the multiphase nature of natural magmas, the validity of either assumption requires future experimental evaluation³.

From our model simulations we find that shear-induced fragmentation is predicted to occur over a wide range of parameters²² encompassing explosive as well as effusive eruption conditions (Fig. 3). The evolution of individual model simulations follows the grey horizontal lines from low to high melt viscosity, until the fragmentation threshold, denoted by dots and squares, is reached (Fig. 3). Our results indicate that the threshold for shear-induced fragmentation follows a systematic trend (diagonal line on Fig. 3)

that scales with the critical conduit shear-strain rate, $\dot{\gamma} = Q/\pi R^3$, via the relaxed melt viscosity as:

$$\dot{\gamma} \approx C G_\infty \mu_{mr}^{-0.9} \quad (6)$$

where $C = 0.01$ (Pa s)^{−0.1} is a fitting parameter. At a given Q and R , shear-induced fragmentation is predicted to occur once μ_{mr} exceeds the threshold value given by equation (6).

The degassed nature of effusive magmas is generally attributed to open system degassing^{7,8,23}, the loss of volatiles from the ascending magma. To what extent degassing is controlled by permeable gas flow laterally into the conduit wall^{1,23}, or vertically through the magma column²⁴, remains an open question. Permeability estimates^{25–27} of $\leq 10^{-13}$ m² imply timescales of the order of weeks to years for permeable gas flow through vesicular magma within the conduit. However, based on recent observations^{10,11} we speculate that shear-induced fragmentation may create, at least temporarily and locally, a magma consisting of individual fragments bound by an interconnected fracture network of high permeability. Degassing of centimetre-size fragments into fractures, concurrent with gas flow through a highly permeable fracture network, could considerably shorten the time required for degassing of magma at and above the initial depth of fragmentation. At the initial occurrence of fragmentation almost all deformation associated with magma ascent is localized over a radial distance, $\delta \approx CR\mu_{mr}^{0.1}$, derived by combining equations (5) and (6) under the assumption of conservation of mass. δ is an estimate for the thickness of fractured, and presumably permeable, magma that is expected to form during one cycle of shear-induced fragmentation. Therefore, regardless of the actual degassing pathway, shear-induced fragmentation may increase the feasibility of open-system degassing.

Observations¹⁰ (Fig. 1) suggest that fragmented magma can

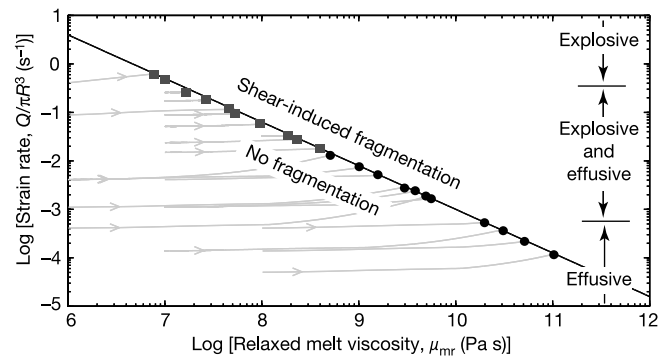


Figure 3 Model results and predicted occurrence of shear-induced fragmentation.

Vertical axis is conduit strain rate (ratio of mean magma ascent velocity and conduit radius). Subhorizontal grey lines with arrows trace the evolution of conduit strain rate and μ_{mr} for individual model simulations. The predicted occurrence of shear-induced fragmentation is marked by symbols: circles, >70% of volatiles originally dissolved in the melt have exsolved at the time of shear-induced fragmentation; squares, an exsolved volatile fraction of <70%. The predicted occurrence of shear-induced fragmentation follows a systematic trend, well described by the power law of equation (6) (solid diagonal line). Below this trend no fragmentation is predicted to occur. A combination of conduit strain rate and μ_{mr} plotting above this trend is inferred to result in shear-induced fragmentation. Also shown are the ranges of explosive versus effusive strain rates for silicic eruptions, based on a compilation of eruption rates²² and assuming an average magma density of approximately 10^3 kg m^{−3}, as well as a minimum (or maximum) conduit radius of the order of 10 m (or 100 m) for explosive (or effusive) eruptions. Our model simulations predict the occurrence of shear-induced fragmentation for both explosive and effusive eruption conditions. The following range of parameters were investigated: $100 \text{ MPa} \leq p_0 \leq 250 \text{ MPa}$, $10^{10} \text{ m}^{-3} \leq n_d \leq 10^{19} \text{ m}^{-3}$, $5 \text{ wt\%} < c(z=0) < 8 \text{ wt\%}$, $\phi(z=0) \approx 0$, $10^5 \text{ Pa s} \leq \mu_m(z=0) \leq 10^8 \text{ Pa s}$, $10^3 \text{ kg s}^{-1} \leq \int_A \rho(z) u(r, z) dA < 10^9 \text{ kg s}^{-1}$, and $5 \text{ m} \leq R \leq 200 \text{ m}$.

reanneal. Although our model simulations do not include calculations past the fragmentation threshold, we propose that a local decrease in shear-strain rates associated with fragmentation may promote reannealing²⁸. Furthermore, it seems reasonable to assume that shear-induced fragmentation has a marked effect on the flow of the ascending magma and that upon continued ascent, fragments from different parts of the ascending magma may become juxtaposed. If the magma is texturally heterogeneous, which in itself may be a consequence of repeated cycles of fragmentation, flow deformation and reannealing, fragments can become elongated into bands¹⁰ (Fig. 1). Minimum strain estimates to produce millimetre-size bands from decimetre-size fragments is of the order of 100. Using δ as an estimate of the length scale for shear, this corresponds to an ascent distance, $\Delta z \approx \gamma_R \delta$, of the order of 10 m. We propose that the long-standing enigma of pervasive flow banding of silicic magmas may in some cases be viewed as a record of fragmentation and reannealing during magma ascent, in much the same way as banding can be made by fragmentation and reannealing in flows²⁹. In addition, we expect that shear-induced fragmentation can, to some degree, replace viscous deformation as the mode of shear along conduit walls, thereby reducing the exceedingly large dynamic pressures required to erupt highly crystalline silicic magmas. However, none of our model simulations explicitly include the effect of crystals on fragmentation³⁰.

Our prediction that shear-induced fragmentation occurs in both explosive and effusive silicic volcanism is consistent with the observed conditions of volcanic systems²² (Fig. 3), with the degassed nature of effusive silicic lavas^{7,8}, and with textural observations at the outcrop scale down to the microscale^{9–11} (Fig. 1). As opposed to the common view that explosive volcanism “is defined as involving fragmentation of magma during ascent”¹, we conclude that fragmentation may play an equally important role in reducing the likelihood of explosive behaviour, by facilitating magma degassing. Because shear-induced fragmentation depends so strongly on the rheology of the ascending magma, our findings are in a broader sense equivalent to Eichelberger’s hypothesis¹ that “higher viscosity of magma may favour non-explosive degassing rather than hinder it”, albeit with the added complexity of shear-induced fragmentation. □

Received 19 May; accepted 15 November 2003; doi:10.1038/nature02138.

- Eichelberger, J. C. Silicic volcanism: ascent of viscous magmas from crustal reservoirs. *Annu. Rev. Earth Planet. Sci.* **23**, 41–63 (1995).
- Dingwell, D. B. Volcanic dilemma: Flow or blow? *Science* **273**, 1054–1055 (1996).
- Papale, P. Strain-induced magma fragmentation in explosive eruptions. *Nature* **397**, 425–428 (1999).
- Dingwell, D. B. & Webb, S. L. Structural relaxation in silicate melts and non-Newtonian melt rheology in geologic processes. *Phys. Chem. Miner.* **16**, 508–516 (1989).
- Webb, S. L. & Dingwell, D. B. The onset of non-Newtonian rheology of silicate melts. *Phys. Chem. Miner.* **17**, 125–132 (1990).
- Webb, S. L. & Dingwell, D. B. Non-Newtonian rheology of igneous melts at high stresses and strain rates: experimental results for rhyolite, andesite, basalt, and nephelinite. *J. Geophys. Res.* **95**, 15695–15701 (1990).
- Newman, S., Epstein, S. & Stolper, E. Water, carbon dioxide and hydrogen isotopes in glasses from the ca. 1340 A.D. eruption of the Mono Craters, California: Constraints on degassing phenomena and initial volatile content. *J. Volcanol. Geotherm. Res.* **35**, 75–96 (1988).
- Villemant, B. & Boudon, G. Transition from dome-forming to plinian eruptive styles controlled by H₂O and Cl degassing. *Nature* **392**, 65–69 (1998).
- Polacci, M., Papale, P. & Rosi, M. Textural heterogeneities in pumices from the climatic eruption of Mount Pinatubo, 15 June 1991, and implications for magma ascent dynamics. *Bull. Volcanol.* **63**, 83–97 (2001).
- Tuffen, H., Dingwell, D. B. & Pinkerton, H. Repeated fracture and healing of silicic magma generates flow banding and earthquakes? *Geology* **31**, 1089–1092 (2003).
- Stasiuk, M. V. et al. Degassing during magma ascent in the Mule Creek vent (USA). *Bull. Volcanol.* **58**, 117–130 (1996).
- Goto, A. A new model for volcanic earthquake at Unzen Volcano: Melt rupture model. *Geophys. Res. Lett.* **26**, 2541–2544 (1999).
- Mastin, L. G. Insights into volcanic conduit flow from an open-source numerical model. *Geochim. Geophys. Geosyst.* **3**, doi:10.1029/2001GC000192 (2002).
- Proussevitch, A. A., Sahagian, D. L. & Anderson, A. T. Dynamics of diffusive bubble growth in magmas: Isothermal case. *J. Geophys. Res.* **3**, 22283–22307 (1993).
- Lensky, N. G., Lyakhovsky, V. & Navon, O. Radial variations of melt viscosity around growing bubbles and gas overpressure in vesiculating magmas. *Earth Planet. Sci. Lett.* **186**, 1–6 (2001).
- Rust, A. C. & Manga, M. Effects of bubble deformation on the viscosity of dilute suspensions. *J. Non-Newtonian Fluid Mech.* **104**, 53–63 (2002).

- Pal, R. Rheological behavior of bubble-bearing magmas. *Earth Planet. Sci. Lett.* **207**, 165–179 (2003).
- Llewellin, E. W., Mader, H. M. & Wilson, S. D. R. The constitutive equation and flow dynamics of bubbly magmas. *Geophys. Res. Lett.* **29**, doi:10.1029/2002GL015697 (2002).
- Simmons, J. H., Mohr, R. K. & Montrose, C. J. Non-Newtonian viscous flow in glass. *J. Appl. Phys.* **53**, 4075–4080 (1982).
- Hess, K.-U. & Dingwell, D. B. Viscosities of hydrous leucogranitic melts: A non-Arrhenian model. *Am. Mineral.* **81**, 1297–1300 (1996).
- Manga, M. & Loewenberg, M. Viscosity of magmas containing highly deformable bubbles. *J. Volcanol. Geotherm. Res.* **105**, 19–24 (2001).
- Pyle, D. M. in *Encyclopedia of Volcanoes* (eds Sigurdsson, H., Houghton, B. F., McNutt, S. R., Rymer, H. & Stix, J.) 263–269 (Academic, San Diego, 2000).
- Jupart, C. & Allegre, C. J. Gas content, eruption rate and instabilities of eruption regime in silicic volcanoes. *Earth Planet. Sci. Lett.* **102**, 413–429 (1991).
- Boudon, G., Villemant, B., Komorowski, J.-C., Ildefonse, P. & Semet, M. P. The hydrothermal system at Soufriere Hills volcano, Montserrat (West Indies): characterization and role in the on-going eruption. *Geophys. Res. Lett.* **25**, 3693–3696 (1998).
- Blower, J. D. Factors controlling porosity-permeability relationships in magma. *Bull. Volcanol.* **63**, 497–504 (2001).
- Klug, C. & Cashman, K. V. Permeability development in vesiculating magmas: implications for fragmentation. *Bull. Volcanol.* **58**, 87–100 (1996).
- Klug, C., Cashman, K. V. & Bacon, C. R. Structure and physical characteristics of pumice from the climatic eruption of Mount Mazama (Crater Lake), Oregon. *Bull. Volcanol.* **64**, 486–501 (2002).
- Gottsmann, J. & Dingwell, D. B. The thermal history of a spatter-fed lava flow: the 8-ka pantellerite flow of Mayor Island, New Zealand. *Bull. Volcanol.* **64**, 410–422 (2002).
- Smith, J. V. Ductile-brittle transition structures in the basal shear zone of a rhyolite lava flow, eastern Australia. *J. Volcanol. Geotherm. Res.* **72**, 217–223 (1996).
- Martel, C., Dingwell, D. B., Spieler, O., Pichavant, M. & Wilke, M. Experimental fragmentation of crystal- and vesicle-bearing melts. *Bull. Volcanol.* **63**, 398–405 (2001).

Acknowledgements We thank P. Papale and D. L. Sahagian for comments on the previous versions of the manuscript, and K. V. Cashman, A. Rust, and A. M. Jellinek for comments on earlier versions. This work was supported by the National Science Foundation and the Sloan Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.M.G. (hmg@seismo.berkeley.edu).

Language-tree divergence times support the Anatolian theory of Indo-European origin

Russell D. Gray & Quentin D. Atkinson

Department of Psychology, University of Auckland, Private Bag 92019, Auckland 1020, New Zealand

Languages, like genes, provide vital clues about human history^{1,2}. The origin of the Indo-European language family is “the most intensively studied, yet still most recalcitrant, problem of historical linguistics”³. Numerous genetic studies of Indo-European origins have also produced inconclusive results^{4,5,6}. Here we analyse linguistic data using computational methods derived from evolutionary biology. We test two theories of Indo-European origin: the ‘Kurgan expansion’ and the ‘Anatolian farming’ hypotheses. The Kurgan theory centres on possible archaeological evidence for an expansion into Europe and the Near East by Kurgan horsemen beginning in the sixth millennium BP^{7,8}. In contrast, the Anatolian theory claims that Indo-European languages expanded with the spread of agriculture from Anatolia around 8,000–9,500 years BP⁹. In striking agreement with the Anatolian hypothesis, our analysis of a matrix of 87 languages with 2,449 lexical items produced an estimated age range for the initial Indo-European divergence of between 7,800 and 9,800 years BP. These results were robust to changes in coding procedures, calibration points, rooting of the trees and priors in the bayesian analysis.

Historical linguists traditionally use the ‘comparative method’ to construct language family trees from discrete lexical, morphological and phonological data. Unfortunately, although the comparative method can provide a relative chronology, it cannot provide absolute date estimates. An alternative method of analysis is glottochronology. This derivative of lexicostatistics is a distance-based approach to language-tree construction that enables absolute dates to be estimated¹⁰. Glottochronology uses the percentage of shared ‘cognates’ between languages to calculate divergence times by assuming a constant rate of lexical replacement or ‘glottoclock’. Cognates are words inferred to have a common historical origin because of systematic sound correspondences and clear similarities in form and meaning. Despite some initial enthusiasm, the method has been heavily criticized and is now largely discredited^{11,12}. Criticisms of glottochronology, and distance-based methods in general, tend to fall into four main categories: first, by summarizing cognate data into percentage scores, much of the information in the discrete character data is lost, greatly reducing the power of the method to reconstruct evolutionary history accurately¹³; second, the clustering methods used tend to produce inaccurate trees when lineages evolve at different rates, grouping together languages that evolve slowly rather than languages that share a recent common ancestor^{12,14}; third, substantial borrowing of lexical items between languages makes tree-based methods inappropriate; and fourth, the assumption of a strict glottoclock rarely holds, making date estimates unreliable¹¹. For these reasons, historical linguists have generally abandoned efforts to estimate absolute ages. Dixon¹⁵ epitomizes this view with his assertion that, on the basis of linguistic data, the age of Indo-European “could be anything—4,000 years BP or 40,000 years BP are both perfectly possible (as is any date in between)”.

Recent advances in computational phylogenetic methods, however, provide possible solutions to the four main problems faced by glottochronology. First, the problem of information loss that comes from converting discrete characters into distances can be overcome by analysing the discrete characters themselves to find the optimal tree(s). Second, the accuracy of tree topology and branch-length estimation can be improved by using models of evolution. Maximum-likelihood methods generally outperform distance and parsimony approaches in situations where there are unequal rates of change¹⁴. Moreover, uncertainty in the estimation of tree topology, branch lengths and parameters of the evolutionary model can be estimated using bayesian Markov chain Monte Carlo¹⁶ (MCMC) methods in which the frequency distribution of the sample approximates the posterior probability distribution of the trees¹⁷. All subsequent analyses can then incorporate this uncertainty. Third, lexical items that are obvious borrowings can be removed from the analysis, and computational methods such as split decomposition¹⁸, which do not force the data to fit a tree model, can be used to check for non-tree-like signals in the data. Finally, the assumption of a strict clock can be relaxed by using rate-smoothing algorithms to model rate variation across the tree. The penalized-likelihood¹⁹ model allows rate variation between lineages while incorporating a ‘roughness penalty’ that penalizes changes in rate from branch to branch. This smoothes inferred rate variation across the tree so that the age of any node can be estimated even under conditions of rate heterogeneity.

We applied likelihood models of lexical evolution, bayesian inference of phylogeny and rate-smoothing algorithms to a matrix of 87 Indo-European languages with 2,449 cognate sets coded as discrete binary characters. This coding was based on the Indo-European database of Dyen *et al.*²⁰, with the addition of three extinct languages. Examining subsets of languages using split decomposition revealed a strong tree-like signal in the data, and a preliminary parsimony analysis produced a consistency index of 0.48 and a retention index of 0.76, well above what would be expected from biological data sets of a similar size²¹. The consensus tree from an

initial analysis is shown in Fig. 1a. The topology of the tree is consistent with the traditional Indo-European language groups²². All of these groups are monophyletic and supported by high posterior probability values. Recent parsimony and compatibility analyses have also supported these groupings, as well as a Romano-Germano-Celtic supergroup, the early divergence of Greek and Armenian lineages²³, and the basal position of Tocharian²⁴. The consensus tree also reflects traditional uncertainties in the relationships between the major Indo-European language groups. For instance, historical linguists have not resolved the position of the Albanian group and our results clearly reflect this uncertainty (the posterior probability of the Albanian/Indo-Iranian group is only 0.36).

One important advantage of the bayesian MCMC approach is that any inferences are not contingent on a specific tree topology. Trees are sampled in proportion to their posterior probability, providing a direct measure of uncertainty in the tree topology and branch-length estimates. By estimating divergence times across the MCMC sample distribution of trees, we can explicitly account for variability in the age estimates due to phylogenetic uncertainty, and hence calculate a confidence interval for the age of any node. We estimated divergence times by constraining the age of 14 nodes on each tree in accordance with historically attested events (see Supplementary Information). We then used penalized-likelihood rate smoothing to calculate divergence times without the assumption of rate constancy¹⁹. Another advantage of the bayesian framework is that prior knowledge of language relationships can be incorporated into the analysis. To ensure that the sample was consistent with well-established linguistic relationships, we filtered the 10,000-tree sample using a constraint tree (Fig. 1b). We used the resulting distribution of 3,500 estimates of basal divergence times to create a confidence interval for the age of the Indo-European language family (Fig. 1b).

A key part of any bayesian phylogenetic analysis is an assessment of the robustness of the inferences. One important potential cause of error is cognacy judgements. In the initial analysis, we included all cognate sets in the Dyen *et al.* database²⁰ in an effort to maximize phylogenetic signal. To assess the impact of different levels of stringency in the cognacy judgements, we repeated the analysis after removing all cognate sets identified by Dyen *et al.* as ‘doubtful’. ‘Doubtful cognates’ (for instance, possible chance similarities) could falsely increase similarities between languages and thus lead to an underestimate of the divergence times. Unrecognized borrowing between closely related languages would have a similar effect. Conversely, borrowing between distantly related languages will falsely inflate branch lengths at the base of the tree and thus increase divergence-time estimates. With the doubtful cognates removed, the conservative coding led to a similar estimate of Indo-European language relationships to that produced using the original coding. The relationships within each of the 11 main groups were unchanged. Only the placement of the weakly supported basal branches differed (Fig. 1c). More significantly, the divergence-time estimates increased, suggesting that the effects of chance similarities and unrecognized borrowings between closely related languages might have outweighed those of borrowings between distantly related languages. In other words, our initial analysis is likely to have underestimated the age of Indo-European.

The constraint tree used to filter the MCMC sample of trees also contained assumptions about Indo-European history that might have biased the results. We therefore repeated the analyses using a more relaxed set of constraints (Fig. 1d). This produced a divergence-time distribution and consensus tree almost identical to the original sample distribution (Fig. 1d).

Another potential bias lay in the initial coding procedure, which made no allowance for missing cognate information. The languages at the base of the tree (Hittite, Tocharian A and Tocharian B) may appear to lack cognates found in other languages because our

knowledge of these extinct languages is limited to reconstructions from ancient texts. This uneven sampling might have increased basal branch lengths and thus inflated estimates of divergence times. We tested this possibility by recoding apparently absent cognates as uncertainties (absent or present) and re-running the analyses.

Although divergence-time estimates decreased slightly, the effect was only small (Fig. 1e).

Finally, although there is considerable support for Hittite (an extinct Anatolian language) as the most appropriate root for Indo-European^{22,23}, rooting the tree with Hittite could be claimed to

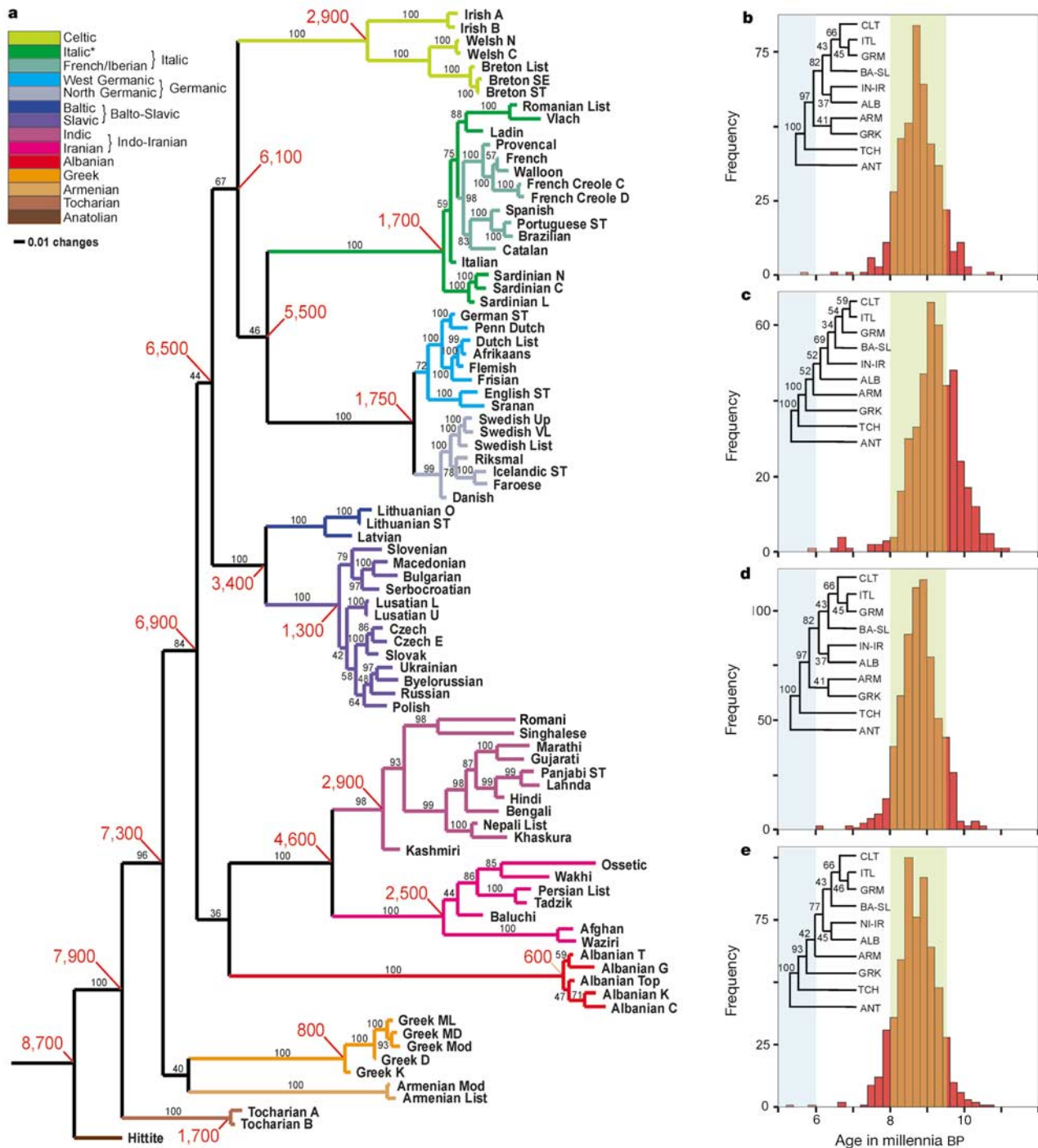


Figure 1 Consensus tree and divergence-time estimates. **a**, Majority-rule consensus tree based on the MCMC sample of 1,000 trees. The main language groupings are colour coded. Branch lengths are proportional to the inferred maximum-likelihood estimates of evolutionary change per cognate. Values above each branch (in black) express the bayesian posterior probabilities as a percentage. Values in red show the inferred ages of nodes in years BP. *Italic also includes the French/Iberian subgroup. Panels **b–e** show the distribution of divergence-time estimates at the root of the Indo-European phylogeny for: **b**, initial assumption set using all cognate information and most stringent constraints (Anatolian, Tocharian, (Greek, Armenian, Albanian, (Iranian, Indic), (Slavic, Baltic), ((North

Germanic, West Germanic), Italic, Celtic)); **c**, conservative cognate coding with doubtful cognates excluded; **d**, all cognate sets with minimum topological constraints (Anatolian, Tocharian, (Greek, Armenian, Albanian, (Iranian, Indic), (Slavic, Baltic), (North Germanic, West Germanic), Italic, Celtic)); **e**, missing data coding with minimum topological constraints and all cognate sets. Shaded bars represent the implied age ranges under the two competing theories of Indo-European origin: blue, Kurgan hypothesis; green, Anatolian farming hypothesis. The relationship between the main language groups in the consensus tree for each analysis is also shown, along with posterior probability values.

bias the analysis in favour of the Anatolian hypothesis. We thus re-ran the analysis using the consensus tree in Fig. 1 rooted with Balto-Slavic, Greek and Indo-Iranian as outgroups. This increased the estimated divergence time from 8,700 years BP to 9,600, 9,400 and 10,100 years BP, respectively.

The pattern and timing of expansion suggested by the four analyses in Fig. 1 is consistent with the Anatolian farming theory of Indo-European origin. Radiocarbon analysis of the earliest Neolithic sites across Europe suggests that agriculture arrived in Greece at some time during the ninth millennium BP and had reached as far as Scotland by 5,500 years BP²⁵. Figure 1 shows the Hittite lineage diverging from Proto-Indo-European around 8,700 years BP, perhaps reflecting the initial migration out of Anatolia. Tocharian, and the Greco-Armenian lineages are shown as distinct by 7,000 years BP, with all other major groups formed by 5,000 years BP. This scenario is consistent with recent genetic studies supporting a Neolithic, Near Eastern contribution to the European gene pool^{14,6}. The consensus tree also shows evidence of a period of rapid divergence giving rise to the Italic, Celtic, Balto-Slavic and perhaps Indo-Iranian families that is intriguingly close to the time suggested for a possible Kurgan expansion. Thus, as observed by Cavalli-Sforza *et al.*²⁶, these hypotheses need not be mutually exclusive.

Phylogenetic methods have revolutionized evolutionary biology over the past 20 years and are now starting to take hold in other areas of historical inference^{2,23,24,27,28,29}. The model-based bayesian framework used in this paper offers several advantages over previous applications of computational methods to language phylogenies. This approach allowed us to: identify sections in the language tree that were poorly supported; explicitly incorporate this uncertainty in tree typology and branch-length estimates in our analysis; test the possible effects of borrowing, chance similarities and bayesian priors on our analysis; and estimate divergence times without the assumption of a strict glottoclock. The challenge of making accurate inferences about human history is an extremely demanding one, requiring the integration of archaeological, genetic, cultural and linguistic data. The combination of computational phylogenetic methods and lexical data to test archaeological hypotheses is a step forward in this challenging and fascinating task. □

Methods

Data and coding

Data were sourced from the comparative Indo-European database created by Dyen *et al.*²⁰. The database records word forms and cognacy judgements in 95 languages across the 200 items in the Swadesh word list. This list consists of items of basic vocabulary such as pronouns, numerals and body parts that are known to be relatively resistant to borrowing. For example, although English is a Germanic language, it has borrowed around 50% of its total lexicon from French and Latin. However, only about 5% of English entries in the Swadesh 200-word list are clear Romance language borrowings¹. Where borrowings were obvious, Dyen *et al.* did not score them as cognate and thus they were excluded from our analysis; 11 of the speech varieties that were not coded by Dyen *et al.* were also excluded. To facilitate reconstruction of some of the oldest language relationships, we added three extinct Indo-European languages, thought to fit near the base of the tree (Hittite, Tocharian A and Tocharian B). Word form and cognacy judgements for all three languages were made on the basis of multiple sources to ensure reliability. The presence or absence of words from each cognate set was coded as '1' or '0', respectively, to produce a binary matrix of 2,449 cognates in 87 languages.

Tree construction

Language trees were constructed using a 'restriction site' model of evolution that allows unequal character-state frequencies and gamma-distributed character-specific rate heterogeneity (MrBayes version 2.01; ref. 30). We used default 'flat' priors for the rate matrix, branch lengths, gamma shape parameter and site-specific rates. The results were found to be robust to changes in these priors. For example, repeating the analyses with an exponential branch-length prior produced a 95% confidence interval for the basal divergence time of between 7,100 and 9,200 years BP.

The program was run ten times using four concurrent Markov chains. Each run generated 1,300,000 trees from a random starting phylogeny. On the basis of an autocorrelation analysis, only every 10,000th tree was sampled to ensure that consecutive samples were independent. A 'burn-in' period of 300,000 trees for each run was used to avoid sampling trees before the run had reached convergence. Log-likelihood plots and an examination of the post-burn-in tree topologies showed that the runs had indeed reached convergence by this time. For each analysis a total of 1,000 trees were sampled and rooted

with Hittite. The branch between Hittite and the rest of the tree was split at the root such that half its length was assigned to the Hittite branch and half to the remainder of the tree; divergence-time estimates were found to be robust to threefold alterations of this allocation.

Divergence-time estimates

Eleven nodes corresponding to the points of initial divergence in all of the major language subfamilies were given minimum and/or maximum ages on the basis of known historical information (see Supplementary Information). The ages of all terminal nodes on the tree, representing languages spoken today, were set to zero by default. Hittite and the Tocharian languages were constrained in accordance with estimated ages of the source texts. Relatively broad date ranges were chosen to avoid making disputable, a priori assumptions about Indo-European history. A likelihood ratio test with the extinct languages removed revealed that rates were significantly non-clock-like ($\chi^2 = 787.3$, d.f. = 82, $P < 0.001$). Divergence-time estimates were thus made using the semi-parametric, penalized-likelihood model of rate variation implemented in R8s (version 1.50)¹⁹. The cross-validation procedure was applied to the majority-rule consensus tree (Fig. 1) to determine the optimal value of the rate-smoothing parameter. Step-by-step removal of each of the 14 age constraints on the consensus tree revealed that divergence-time estimates were robust to calibration errors. For 13 nodes, the reconstructed age was within 390 years of the original constraint range. Only the reconstructed age for Hittite showed an appreciable variation from the constraint range. This may be attributable to the effect of missing data associated with extinct languages. Reconstructed ages at the base of the tree ranged from 10,400 years BP with the removal of the Hittite age constraint, to 8,500 years BP with the removal of the Iranian group age constraint.

Received 18 July; accepted 22 August 2003; doi:10.1038/nature02029.

1. Pagel, M. in *Time Depth in Historical Linguistics* (eds Renfrew, C., McMahon, A. & Trask, L.) 189–207 (The McDonald Institute for Archaeological Research, Cambridge, UK, 2000).
2. Gray, R. D. & Jordan, F. M. Language trees support the express-train sequence of Austronesian expansion. *Nature* **405**, 1052–1055 (2000).
3. Diamond, J. & Bellwood, P. Farmers and their languages: the first expansions. *Science* **300**, 597–603 (2003).
4. Richards, M. *et al.* Tracing European founder lineage in the Near Eastern mtDNA pool. *Am. J. Hum. Genet.* **67**, 1251–1276 (2000).
5. Semoni, O. *et al.* The genetic legacy of Paleolithic *Homo sapiens* in extant Europeans: a Y chromosome perspective. *Science* **290**, 1155–1159 (2000).
6. Chikhi, L., Nichols, R. A., Barbujani, G. & Beaumont, M. A. Y genetic data support the Neolithic Demic Diffusion Model. *Proc. Natl Acad. Sci. USA* **99**, 11008–11013 (2002).
7. Gimbutas, M. The beginning of the Bronze Age in Europe and the Indo-Europeans 3500–2500 B.C. *J. Indo-Eur. Stud.* **1**, 163–214 (1973).
8. Mallory, J. P. *Search of the Indo-Europeans: Languages, Archaeology and Myth* (Thames & Hudson, London, 1989).
9. Renfrew, C. in *Time Depth in Historical Linguistics* (eds Renfrew, C., McMahon, A. & Trask, L.) 413–439 (The McDonald Institute for Archaeological Research, Cambridge, UK, 2000).
10. Swadesh, M. Lexico-statistic dating of prehistoric ethnic contacts. *Proc. Am. Phil. Soc.* **96**, 453–463 (1952).
11. Bergsland, K. & Vogt, H. On the validity of glottochronology. *Curr. Anthropol.* **3**, 115–153 (1962).
12. Blust, R. in *Time Depth in Historical Linguistics* (eds Renfrew, C., McMahon, A. & Trask, L.) 311–332 (The McDonald Institute for Archaeological Research, Cambridge, UK, 2000).
13. Steel, M. A., Hendy, M. D. & Penny, D. Loss of information in genetic distances. *Nature* **333**, 494–495 (1988).
14. Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. in *Molecular Systematics* (eds Hillis, D., Moritz, C. & Mable, B. K.) 407–514 (Sinauer Associates, Inc., Sunderland, Massachusetts, 1996).
15. Dixon, R. M. W. *The Rise and Fall of Language* (Cambridge Univ. Press, Cambridge, UK, 1997).
16. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1091 (1953).
17. Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310–2314 (2001).
18. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).
19. Sanderson, M. R8s. *Analysis of Rates of Evolution, Version 1.50* (Univ. California, Davis, 2002).
20. Dyen, I., Kruskal, J. B. & Black, P. *FILE IE-DATA1*. Available at (<http://www.ntu.edu.au/education/langs/ielex/IE-DATA1>) (1997).
21. Sanderson, M. J. & Donoghue, M. J. Patterns of variation in levels of homoplasy. *Evolution* **43**, 1781–1795 (1989).
22. Gamkrelidze, T. V. & Ivanov, V. V. *Trends in Linguistics 80: Indo-European and the Indo-Europeans* (Mouton de Gruyter, Berlin, 1995).
23. Rexova, K., Frynta, D. & Zrzavy, J. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* **19**, 120–127 (2003).
24. Ringe, D., Warnow, T. & Taylor, A. Indo-European and computational cladistics. *Trans. Philol. Soc.* **100**, 59–129 (2002).
25. Gkiasta, M., Russell, T., Shennan, S. & Steele, J. Neolithic transition in Europe: the radiocarbon record revisited. *Antiquity* **77**, 45–62 (2003).
26. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, 1994).
27. Holden, C. J. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc. R. Soc. Lond. B* **269**, 793–799 (2002).
28. Barbrook, A. C., Howe, C. J., Blake, N. & Robinson, P. The phylogeny of The Canterbury Tales. *Nature* **394**, 839 (1998).
29. McMahon, A. & McMahon, R. Finding families: Quantitative methods in language classification. *Trans. Philol. Soc.* **101**, 7–55 (2003).
30. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* **17**, 754–755 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank S. Allan, L. Campbell, L. Chikhi, M. Corballis, N. Gavey, S. Greenhill, J. Hamm, J. Huelsenbeck, G. Nichols, A. Rodrigo, F. Ronquist, M. Sanderson and S. Shennan for useful advice and/or comments on the manuscript.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.G. (rd.gray@auckland.ac.nz).

Contributions of microbial biofilms to ecosystem processes in stream mesocosms

Tom J. Battin¹, Louis A. Kaplan², J. Denis Newbold² & Claude M. E. Hansen³

¹Department of Limnology, IECB, University of Vienna, A-1090 Vienna, Austria

²Stroud Water Research Center, Avondale 19311 Philadelphia, USA

³Institute of Limnology and Zoology, University of Innsbruck, A-6020 Innsbruck, Austria

In many aquatic ecosystems, most microbes live in matrix-enclosed biofilms^{1–3} and contribute substantially to energy flow and nutrient cycling. Little is known, however, about the coupling of structure and dynamics of these biofilms to ecosystem function². Here we show that microbial biofilms changed the physical and chemical microhabitat and contributed to ecosystem processes in 30-m-long stream mesocosms. Biofilm growth increased hydrodynamic transient storage—streamwater detained in quiescent zones, which is a major physical template for ecological processes in streams^{4,5}—by 300% and the retention of suspended particles by 120%. In addition, by enhancing the relative uptake of organic molecules of lower bioavailability, the interplay of biofilm microarchitecture and mass transfer changed their downstream linkage. As living zones of transient storage, biofilms bring hydrodynamic retention and biochemical processing into close spatial proximity and influence biogeochemical processes and patterns in streams. Thus, biofilms are highly efficient and successful ecological communities that may also contribute to the influence that headwater streams have on rivers, estuaries and even oceans^{6,7} through longitudinal linkages of local biogeochemical and hydrodynamic processes.

Although the broad physical factors that influence ecological processes at the streambed interface have been extensively studied^{4,5,8}, the interactions of underlying biological, chemical and physical mechanisms operating at the microscale have not. To investigate whether biofilm growth changes hydrodynamic transient storage and organic matter processing at the streambed/streamwater interface we experimented with natural microbial biofilms in duplicate streamside flumes under two open-channel flow velocities (Methods and Supplementary Information). In both flow treatments, initial biofilms consisted of largely bacterial microcolonies that rapidly coalesced into a basal geometric film surrounding conspicuous voids (Fig. 1). Diatoms became a major component in mature biofilms, and long streamers (filamentous structures) oscillating in the water flow developed in these biofilms. Flow significantly affected biofilm development, yielding higher biomass under slower flows. Biofilm detachment and invertebrate grazing dramatically decreased biomass after day 14 in fast flow and day 25 in slow flow. Confocal microscope analyses of biofilm cryosections (x – z plane) revealed that flow also shaped biofilm microarchitecture beyond bulk biomass. Biofilms grown in slow flow developed clearly

visible skins of diatoms, and were thicker with higher surface sinuosity and elevated density than biofilms grown in the fast-flow treatment.

Transient storage has typically been associated with physical structures such as the interstices within streambed sediments or quiescent waters in back eddies, pools and side channels^{4,5}. Although others^{9,10} have indicated the potential for algal-dominated biofilms to influence stream hydrodynamics, we have quantified biofilms as a significant component of the transient storage zone in the stream mesocosms. To do so, we estimated A_s , the cross-sectional area of transient storage (in m^2 , see Methods) and compared it to the cross-sectional areas of the total biofilm, including the biofilm voids. At day 0 (that is, no biofilms) A_s was entirely due to the flume bed, and differences between treatments were attributable to different flow velocities. As microbial growth progressed, transient storage increased 4- and 2.3-fold in the slow- and fast-flow treatments, respectively (Fig. 2). This increase closely followed the temporal pattern of biofilm dynamics, with chlorophyll a and organic matter together explaining 76% of the variance in A_s (multiple linear regression, MLR: degrees of freedom,

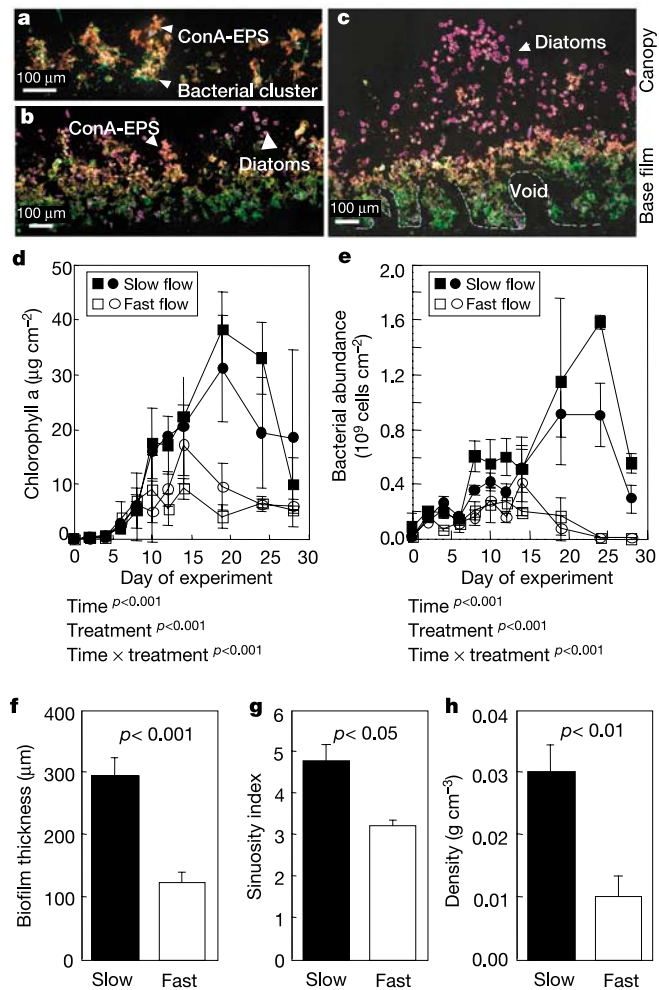


Figure 1 Effects of flow on biofilm structure and microarchitecture. **a–c**, Confocal scanning laser micrographs of 3-day-old (**a**), 15-day-old (**b**) and 24-day-old (**c**) biofilms. ConA-EPS, concanavalin-A Texas Red stained exopolysaccharides. **d** and **e**, dynamics of biofilm associated chlorophyll a (**d**) and bacterial abundance (**e**). Given are mean $\pm$ s.d. of two or three independent measurements. Circles and squares in all figures denote separate replicate mesocosms. **f–h**, Elevated biofilm thickness (**f**), surface sinuosity (**g**) and density (**h**) in slow-flow mesocosms. Given are mean $\pm$ s.d. of measurements made on seven to ten different dates for each flow treatment.

d.f. = 22, $p < 0.01$). Unlike A_s , the hydraulic exchange velocity—the piston flow into transient storage—was higher in the fast flows. Yet within each treatment, it followed a temporal pattern similar to that of A_s . Cross-sectional areas (m^2) of the total biofilm ($r^2 = 0.81$, d.f. = 10, $p < 0.001$) and of biofilm voids as derived from porosity ($r^2 = 0.89$, d.f. = 10, $p < 0.001$) correlated with the added transient storage zone (m^2) (that is, A_s corrected for the storage zone generated by the flume bed at day 0). Total biofilm and biofilm void cross-sectional areas were respectively, 33% and 16% as large as the added transient storage zone. This indicates that both the internal biofilm channel system (see voids, Fig. 1c) and the highly hydrated³ matrix constitute important transient storage zones—and that the increased transient storage not only contains the growing biofilms, but extends beyond them. The relationship ($r^2 = 0.56$, d.f. = 10, $p < 0.01$) between transient storage and biofilm surface sinuosity suggests how microarchitecture increases transient storage around biofilms. Protuberances at the biofilm surface and streamers could generate vortices that create additional transient storage, which in turn influences the delivery of solutes from the water column to the biofilms. If the oscillation of streamers actually enhances mass transfer within biofilms¹¹, it is likely that oscillating streamers from biofilms covering an entire streambed would collectively increase hydrodynamic exchange at the streambed interface. This underscores the ecological significance of the microbial-altered transient storage.

As biofilm growth altered the hydrodynamic environment in the flumes, it simultaneously increased the deposition velocity of suspended organic particles (Fig. 3a). Particle deposition was predicted best from the hydraulic exchange velocity and accumulation of microbial exopolysaccharides (MLR: $r^2 = 0.78$, d.f. = 16, $p < 0.001$). Hydraulic exchange velocity has been shown to influence inorganic, physicochemical particle deposition in streams⁸, but

our studies found that the hydraulic exchange velocity was five- to eightfold lower than the particle deposition velocity (compare Figs 2b and 3a). This indicates that advective delivery could not have accounted for most of the deposition. Thus, the relationship between microbial growth and deposition provides evidence that biofilm stickiness³ is an important biological influence on particle deposition in streams. This may account for hitherto unexplained variability in measurements of particle deposition^{12,13}. Furthermore, there was a positive relationship ($r^2 = 0.72$, d.f. = 7, $p = 0.01$) between biofilm microarchitecture (as sinuosity) and particle deposition. Our finding that complex surfaces with small streamers or mushroom-shaped features enhance particle retention agrees with the few existing results from laboratory-grown bacterial biofilms¹⁴.

Biofilm theory^{15,16} holds that solute retention, or the uptake of organic substrates as a source of carbon and energy, is diffusion limited. Before assimilation, solutes must pass first from the bulk liquid to the biofilm surface (external mass transfer) and then through the biofilm matrix to the cells (internal mass transfer). Consequently, as biofilms grow, mass transfer may become the primary rate-limiting step because substrate consumption and biofilm structure increase resistance to transport. To determine how biofilm growth influences solute uptake, we measured mass transfer coefficients (V) for the uptake from the water column of glucose and arabinose—two monomeric carbohydrates differing in bioavailability¹⁷—which we had experimentally injected into the flumes. Average uptake of glucose was significantly higher than arabinose in both the slow (2.5-fold, $p < 0.001$, d.f. = 26) and fast (2.3-fold, $p < 0.001$, d.f. = 26) flow treatment, and both monomers showed trends of increasing uptake as biofilms grew (Supplementary Information). However, as biofilms grew, both the absolute and relative uptake rates of arabinose increased, that is

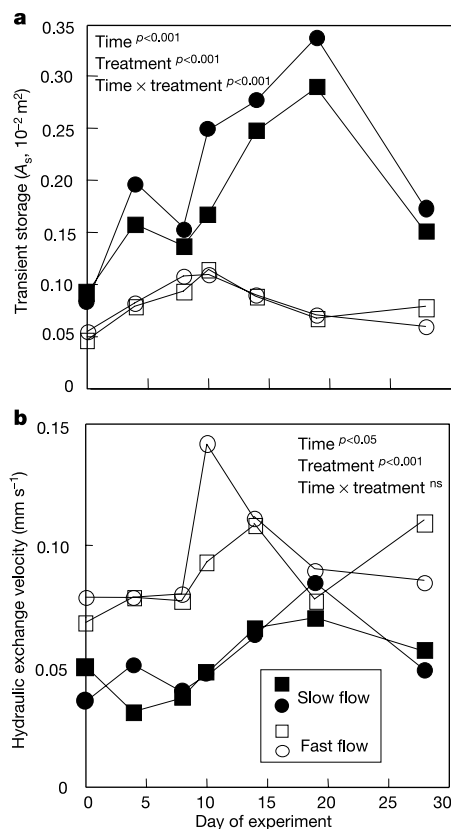


Figure 2 Temporal dynamics of the transient storage zone as A_s (a) and the vertical exchange velocity between the water column and the transient storage (b).

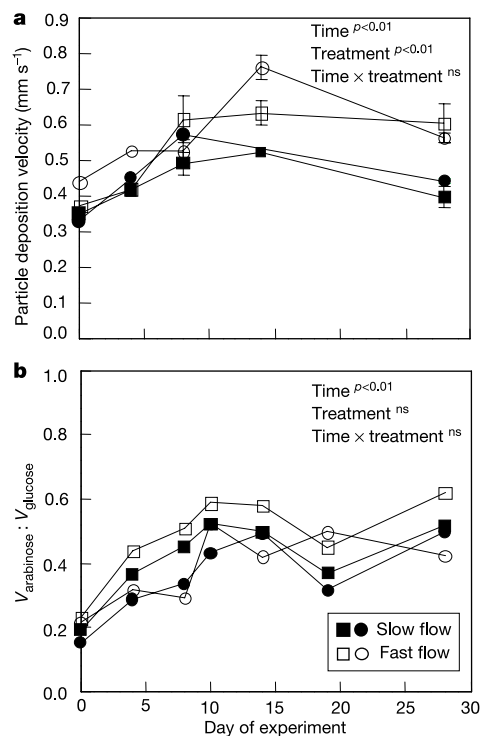


Figure 3 Temporal dynamics of the deposition velocity of suspended organic particles (a) and of the ratio between the mass transfer coefficients for arabinose and glucose (b). Error bars in a indicate the s.d. of the mean of two to five independent particle injections.

$V_{\text{arabinose}}:V_{\text{glucose}}$ increased (Fig. 3b). Thus, biofilm development affected mass transfer, which influenced the relative uptake of substrates differing in bioavailability. These data illustrate the general phenomenon that as transport of solutes from the water column into the biofilm becomes limiting, the susceptibility of a molecule to metabolism shifts from its intrinsic biological lability to the ability to diffuse into the biofilm. Depending on the developmental stage therefore, biofilms reduce the differences in the apparent bioavailability of substrates and compress the longitudinal coupling of substrate production and consumption to shorter distances within a stream reach.

Underlying these mass transfer processes and uptake kinetics is the physical structure of a biofilm¹. Recent work^{1,11,18} has contradicted the traditional view of biofilms as homogeneous structures subject to diffusion limitation of external and internal mass transport by demonstrating their heterogeneous architecture with channel networks, mushroom-shaped protuberances and oscillating streamers. Such structural heterogeneity presents modelling challenges and suggests that some internal transport may be advective^{1,18}. Our experimental measurements of uptake ratios proved useful in assessing the influence of complex structures on transport and uptake kinetics. We found that the $V_{\text{arabinose}}:V_{\text{glucose}}$ ratio not only increased with bacterial abundance (log-phase: $r^2 = 0.90$, d.f. = 16, $p < 0.001$) and biofilm density ($r^2 = 0.67$, d.f. = 7, $p < 0.01$), parameters that should clearly induce transport limitation, but also decreased with biofilm porosity ($r^2 = 0.64$, d.f. = 8, $p < 0.01$) and fragmentation ($r^2 = 0.41$, d.f. = 8, $p < 0.05$), factors that should enhance mass transport. In fact, elevated porosity translates into a larger channel system that favours advective mass transport¹⁸ whereas highly fragmented biofilm clusters with large surface to volume ratios may decrease the average diffusive path length of a substrate molecule from the bulk liquid to cells within a biofilm cluster¹⁹.

Our results highlight the significance of microscale coupling between biofilm structure and function for stream ecosystem biogeochemistry. This focus complements previous considerations of dynamics operating only at the scale of stream reaches^{4,5,8} and is a prerequisite for successful restoration and preservation practices. Because headwater streams are most vulnerable to human alterations, it is imperative to understand the contributions of microbial biofilms to these ecosystems. □

Methods

Cultivation of biofilms

Biofilms were cultivated in four streamside flume mesocosms (length, 30 m; width, 0.30 m; depth, 0.30 m) adjacent to the White Clay Creek (WCC, Stroud Water Research Center). Flumes were scaled to simulate stream natural dynamics and were continuously fed in a once-through mode with raw stream water. Flume slopes were adjusted to 0.0021 and 0.024 m m⁻¹ yielding open channel flow velocity treatments in duplicate with 0.065 ± 0.003 and 0.23 ± 0.005 m s⁻¹, respectively, which translate into the dimensionless Reynolds numbers of 1,869 ± 190 and 7,559 ± 68, respectively. Flumes were packed with the same clean, biofilm-free sediment (median grain size, 20.9 ± 0.8 mm) containing small, unglazed ceramic tiles for the analysis of microbial parameters. Biofilm parameters were monitored every 2 to 3 days on randomly sampled ceramic tiles during the 30-day experiment.

Microscopy and microbial biomass

Confocal scanning laser microscopy and image analysis (Arc-View 2.2, ESRI) of cryosections were used to describe the micro-architecture of biofilms²⁰. Bacteria and certain exopolysaccharides were stained with SYTO 13 and Concanavalin-A Texas Red, respectively²¹. Biofilm porosity was determined as the ratio of the void surface area to the total cross-sectional surface area of cryosections. Sinuosity of the biofilm surface was determined as the ratio between the curvilinear and the linear distance between two given points of the biofilm front²⁰. Biofilm fragmentation was measured as the ratio between the perimeter and surface area of biofilm clusters. Biofilm density was derived from average thickness and total mass estimates. Bacterial abundance, chlorophyll *a* and exopolysaccharides were assayed as described elsewhere²⁰.

Hydrodynamic exchange

Chloride was injected as a conservative tracer into each flume and we monitored conductivity in flume effluents. Hydrodynamic exchange parameters were estimated by least squares using OTIS-P (ref. 22), an advection-dispersion model that includes transient

storage. Estimated parameters were cross-sectional areas of the transient storage zone A_s (m²) and water column A (m²), flow velocity (m s⁻¹), and the storage zone exchange coefficient α (s⁻¹). The equivalent hydraulic exchange velocity (piston velocity) was calculated as $v = \alpha d$, where d is the average water depth. Biofilm cross-sectional area (m²) in the flumes was calculated from the mean biofilm thickness and the average perimeter of the sediment grains in the mesocosms.

Particle and carbohydrate dynamics

Natural suspended particles were collected from WCC and sieved to retain the 53 to 106 µm fraction. Duplicate or triplicate aliquots were injected into each flume, and the effluent was passed through a 50-µm Nitex net to capture particles in transport. Particle mass was corrected for baseline transport and expressed as dry mass (dried at 100 °C) and ash-free dry mass (4 h, 450 °C).

We injected D-(+)-glucose and D-(-)-arabinose along with chloride into each flume to increase monomeric background concentrations from <10 nM to approximately 200 nM. Injections lasted 20 and 15 min in the slow- and fast-flow treatments, respectively (same as for chloride). Triplicate water samples were collected during the conductivity plateau at an upper station and in the flume effluent, filtered (0.2 µm), and frozen pending analysis by HPLC-PAD²³. Arabinose and glucose concentrations were corrected for background concentration.

The transport distances, S (m), of particles and carbohydrates in the water column before uptake were estimated from $S = L \ln(M_{\text{input}}/M_{\text{output}})$, where L is the flume length and M is the input and output mass, respectively²⁴. We divided the width-specific discharge (equivalent to the product of average velocity and depth) by S to yield the deposition velocity for particles and the mass transfer coefficient (mm s⁻¹) for arabinose and glucose. The mass transfer coefficient corrects for the effects of water depth and velocity on uptake and represents the demand by the streambed microbial community for a given molecule relative to its concentration in the water column.

Data analysis

Two-way analysis of variance (ANOVA) was used to test for effects of flow and time on microbial biomass, hydrodynamics, particle and solute dynamics. Kruskal-Wallis one-way ANOVA was used to test for flow effects on biofilm micro-architectural parameters. Stepwise MLR analysis was used to explore the variation of hydrodynamic exchange parameters and particle deposition velocity. Data are given as mean ± s.d.

Received 12 June; accepted 21 October 2003; doi:10.1038/nature02152.

- Costerton, J. W., Lewandowski, Z., Caldwell, D. E., Korber, D. R. & Lappin-Scott, H. M. Microbial biofilms. *Annu. Rev. Microbiol.* **49**, 711–745 (1995).
- Palmer, R. J. & White, D. C. Developmental biology of biofilms: implications for treatment and control. *Trends Microbiol.* **5**, 435–440 (1997).
- Sutherland, I. W. The biofilm matrix—an immobilized but dynamic microbial environment. *Trends Microbiol.* **9**, 222–227 (2001).
- Jones, J. B. & Mulholland, P. J. *Streams and Ground Waters* (Academic, San Diego, London, 2000).
- Peterson, B. J. *et al.* Control of nitrogen export from watersheds by headwater streams. *Science* **292**, 86–90 (2001).
- Alexander, R. B., Smith, R. A. & Schwarz, G. E. Effect of stream channel size on the delivery of nitrogen to the Gulf of Mexico. *Nature* **403**, 758–761 (2000).
- Gomi, T., Sidle, R. C. & Richardson, J. S. Understanding processes and downstream linkages of headwater systems. *Bioscience* **52**, 905–916 (2002).
- Packman, A. L., Brooks, N. H. & Morgan, J. J. A physicochemical model for colloid exchange between a stream and a sand bed with bed forms. *Wat. Resour. Res.* **36**, 2351–2361 (2000).
- Mulholland, P. J., Steinman, A. D., Marzolf, E. R., Hart, D. R. & DeAngelis, D. L. Effect of periphyton biomass on hydraulic characteristics and nutrient cycling in streams. *Oecologia* **8**, 40–47 (1994).
- Kim, B. K. & Jackman, A. P. Modeling transient storage and nitrate uptake kinetics in a flume containing a natural periphyton community. *Wat. Resour. Res.* **26**, 505–515 (1990).
- Kugaprasatham, S., Nagata, H. & Ohgaki, S. Effect of turbulence on nitrifying biofilms at non-limiting substrate conditions. *Water Res.* **26**, 1629–1638 (1992).
- Minshall, G. W., Thomas, S. A., Newbold, J. D., Monaghan, M. T. & Cushing, C. E. Physical factors influencing fine organic particle transport and deposition in streams. *J. N. Am. Benthol. Soc.* **19**, 1–16 (2000).
- Thomas, S. A. *et al.* The influence of particle size on seston deposition in streams. *Limnol. Oceanogr.* **46**, 1415–1424 (2001).
- Drury, W. J., Characklis, W. G. & Stewart, P. S. Interactions of 1 µm latex particles with *Pseudomonas aeruginosa* biofilms. *Water Res.* **27**, 1119–1126 (1993).
- Grady, C. P. L. Jr, Daigger, G. T. & Lim, H. C. *Biological Wastewater Treatment* 2nd edn (Marcel Dekker, New York, 1999).
- Gantzer, C. J., Rittmann, B. E. & Herricks, E. E. Mass transport of biodegradable materials to streambed biofilms. *Water Res.* **22**, 709–723 (1988).
- Kaplan, L. A. & Newbold, J. D. in *Aquatic Ecosystems Interactivity of Dissolved Organic Matter* (eds Findlay, S. E. G., & Sinsabaugh, R. L.) 97–119 (Academic, San Diego, London, 2003).
- Picioreanu, C., Van Loosdrecht, M. C. M. & Heijnen, J. J. A theoretical study on the effect of surface roughness on mass transport and transformation in biofilms. *Biotechnol. Bioeng.* **68**, 355–369 (2000).
- Yang, X. M., Beyenal, H., Harkin, G. & Lewandowski, Z. Quantifying biofilm structure using image analysis. *J. Microbiol. Methods* **39**, 109–119 (2000).
- Battin, T. J., Kaplan, L. A., Newbold, J. D., Cheng, X. & Hansen, C. Effects of current velocity on the nascent architecture of stream microbial biofilms. *Appl. Environ. Microbiol.* **69**, 5443–5452 (2003).
- Lawrence, J. R., Neu, T. R. & Swerhorne, G. D. H. Application of multiple parameter imaging for the quantification of algae, bacterial, and exopolymer components of microbial biofilms. *J. Microbiol. Methods* **32**, 253–261 (1998).
- Runkle, R. L. *One-Dimensional Transport with Inflow and Storage (Otis): A Solute Transport Model for Streams and Rivers*. Water-Resources Investigations Report 98–4018 (US Geological Survey, Denver, 1998).

23. Cheng, X. & Kaplan, L. A. Improved analysis of dissolved carbohydrates in stream water with HPLC-PAD. *Anal. Chem.* **73**, 458–461 (2001).
24. Newbold, J. D., Elwood, J. W., O'Neill, R. V. & Van Winkle, W. Measuring nutrient spiraling in streams. *Can. J. Fish. Aquat. Sci.* **38**, 860–863 (1981).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank T. Georgian for insights regarding experimental design and the influence of biofilms on particle deposition; S. Roberts, M. Gentile and X. Cheng for technical assistance; K. Czymmek for CLSM support; and A. I. Packman, R. Sommaruga, G. Singer and J. Blaine for critically reading the manuscript. This work was supported by grants from the Austrian FWF (T.J.B.), the US NSF (L.A.K. and J.D.N.) and the Pennswood Fund for Environmental Research (L.A.K. and J.D.N.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.J.B. (tomba@pflaphy.pph.univie.ac.at).

Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex

Michael Wehr & Anthony M. Zador

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA

Neurons in the primary auditory cortex are tuned to the intensity and specific frequencies of sounds, but the synaptic mechanisms underlying this tuning remain uncertain. Inhibition seems to have a functional role in the formation of cortical receptive fields, because stimuli often suppress similar or neighbouring responses^{1–3}, and pharmacological blockade of inhibition broadens tuning curves^{4,5}. Here we use whole-cell recordings *in vivo* to disentangle the roles of excitatory and inhibitory activity in the tone-evoked responses of single neurons in the auditory cortex. The excitatory and inhibitory receptive fields cover almost exactly the same areas, in contrast to the predictions of classical lateral inhibition models. Thus, although inhibition is typically as strong as excitation, it is not necessary to establish tuning, even in the receptive field surround. However, inhibition and excitation occurred in a precise and stereotyped temporal sequence: an initial barrage of excitatory input was rapidly quenched by inhibition, truncating the spiking response within a few (1–4) milliseconds. Balanced inhibition might thus serve to increase the temporal precision⁶ and thereby reduce the randomness of cortical operation, rather than to increase noise as has been proposed previously⁷.

Receptive fields are shaped by the interaction of excitation and inhibition. In the auditory cortex, indirect measurements of this interaction by extracellular methods^{1–5} have suggested that it follows a classic organization—lateral inhibition—in which stimulation in the receptive field centre elicits excitation, and stimulation in the surround elicits inhibition⁸. Here we have used *in vivo* whole-cell patch-clamp methods to test that model directly by measuring synaptic inputs evoked by pure tones in rat auditory cortical neurons⁹ (Fig. 1). For a subset of neurons (31/62), we improved the isolation of synaptic conductances by blocking action potentials with the fast sodium-channel blocker QX-314.

Tone-evoked synaptic conductance changes were large (Fig. 1c), comparable in magnitude to the largest conductance changes elicited by visual stimulation in area V1 neurons^{10–12}. The true

conductances were almost certainly higher than the values presented here, because voltage escape at the subsynaptic membrane due to electrotonic effects (space-clamp errors) causes an underestimation of synaptic conductances (see Supplementary Information). We

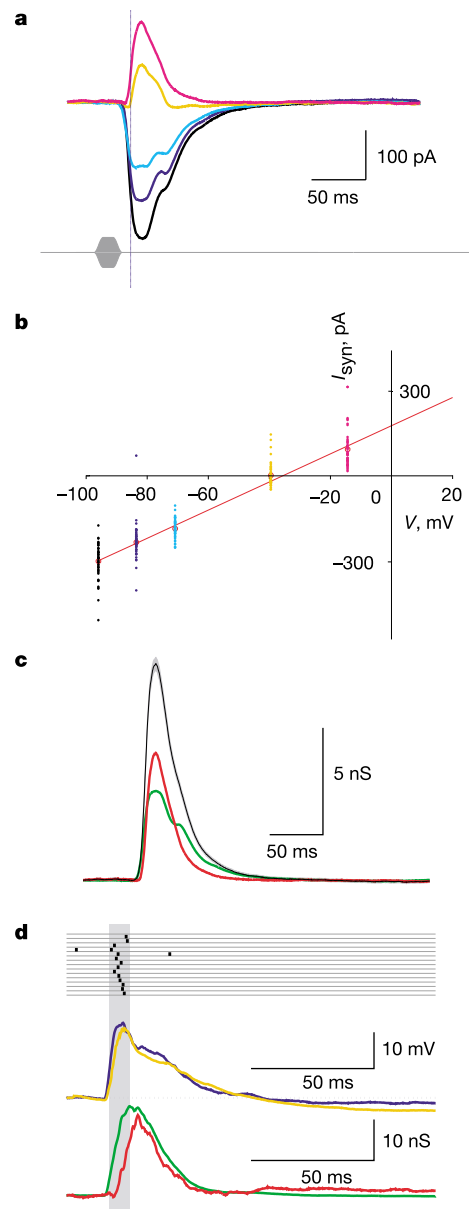


Figure 1 Tones evoked large transient excitatory and inhibitory conductances of comparable magnitudes. **a**, Synaptic currents evoked by a 25-ms tone pip at 1.2 kHz and 66 dB SPL at five holding potentials (48 repetitions). An offset response follows about 25 ms after the onset response. Spikes were blocked with QX-314. **b**, Instantaneous synaptic current-voltage ($I-V$) curve. Synaptic currents (coloured dots) are plotted against holding potential (colours as in **a**) for each of the 48 repetitions (means are indicated by red open circles) at the time of half-maximal synaptic conductance (33 ms from tone onset, indicated by the vertical blue line in **a**). The regression slope and x-intercept give, respectively, the instantaneous synaptic conductance (5.0 nS) and synaptic reversal potential (-36 mV). **c**, Continuous synaptic conductance (black line) with 95% regression confidence limits (shaded region), and its decomposition into excitatory (green) and inhibitory (red) conductances. Across all cells for which spikes were blocked, the maximal conductance evoked in each cell was 7.9 ± 9.7 nS, or $46 \pm 54\%$ of resting input conductance ($n = 31$ cells). **d**, In a different cell, for which spikes were not blocked, excitatory and inhibitory conductances (bottom, green and red traces) predicted the spikes (top, raster plot). The predicted (middle, blue trace, see Supplementary Methods) and actual (middle, orange trace) membrane potentials agreed closely.

decomposed the conductances into their underlying excitatory and inhibitory components^{9,12,13} (see Supplementary Methods). The inhibitory conductance was typically as large as the excitatory conductance, which is consistent with the presence of shunting inhibition¹⁰. Each component followed approximately the same time course as the overall conductance waveform, showing a short-latency rise followed by a slower decay. Excitation evoked a transient spiking response (consisting of typically one or at most a few spikes; Fig. 1d) before being quenched by inhibition after a few milliseconds.

The total synaptic conductance was tuned for sound frequency (Fig. 2). The excitatory and inhibitory components were also tuned; moreover, they were co-tuned, showing approximately the same dependence on stimulus parameters in a given cell (Fig. 2a–d). Tuning of spiking responses was narrower than that of conductances or membrane potential (Fig. 3c), which is consistent with findings in auditory and other cortical regions^{13–16}. Co-tuning differs from classical lateral inhibition⁸, because excitation was at least as broadly tuned as inhibition, and sometimes more so; some surround stimuli elicited a purely excitatory response, whereas pure inhibition was never observed (Fig. 3a). The co-tuning we observed is therefore different from the organization described in some simple cells in primary visual cortex, in which excitation and inhibition can each be elicited in isolation within discrete spatial subregions^{11,12}, but is reminiscent of other simple cells in which excitation and inhibition are evoked together from each sub-region¹⁰.

Because inhibition had the same tuning as excitation, the tuning of the excitatory component alone would, in the absence of timing, be sufficient to establish a neuron's dependence on frequency. Although co-tuned inhibition sharpens tuning, this effect is equivalent to that of any untuned decrease in excitability, such as an increase in spike threshold. By means of this 'iceberg effect,' inhibition (shunting or otherwise) determines how much of the iceberg is suprathreshold but does not affect the shape of the iceberg itself (Fig. 3b); our results are therefore consistent with the finding that pharmacological blockade of inhibition broadens frequency tuning^{4,5}. In this regard, frequency tuning in auditory cortex is similar to orientation tuning in some simple cells in primary visual cortex, which is sharpened by inhibition but which arises mainly through the convergence of thalamocortical inputs^{17,18}. In contrast, orientation selectivity in other visual cortical neurons arises through a diversity of synaptic mechanisms¹³ and can be influenced by recurrent cortical circuitry¹⁹.

Two-tone suppression, in which one tone modifies (usually suppresses) the response to a later tone, has been interpreted as evidence of lateral inhibition in auditory cortical neurons^{1,2,9,20}. Although our observations can explain how tuning is sharpened without lateral inhibition, they cannot by themselves account for two-tone suppression, which can last hundreds of milliseconds—longer than the inhibitory conductances we recorded (but see ref. 9). These suppressive effects might therefore be inherited from pre-synaptic neurons, or they might involve other mechanisms such as short-term synaptic plasticity^{21,22}.

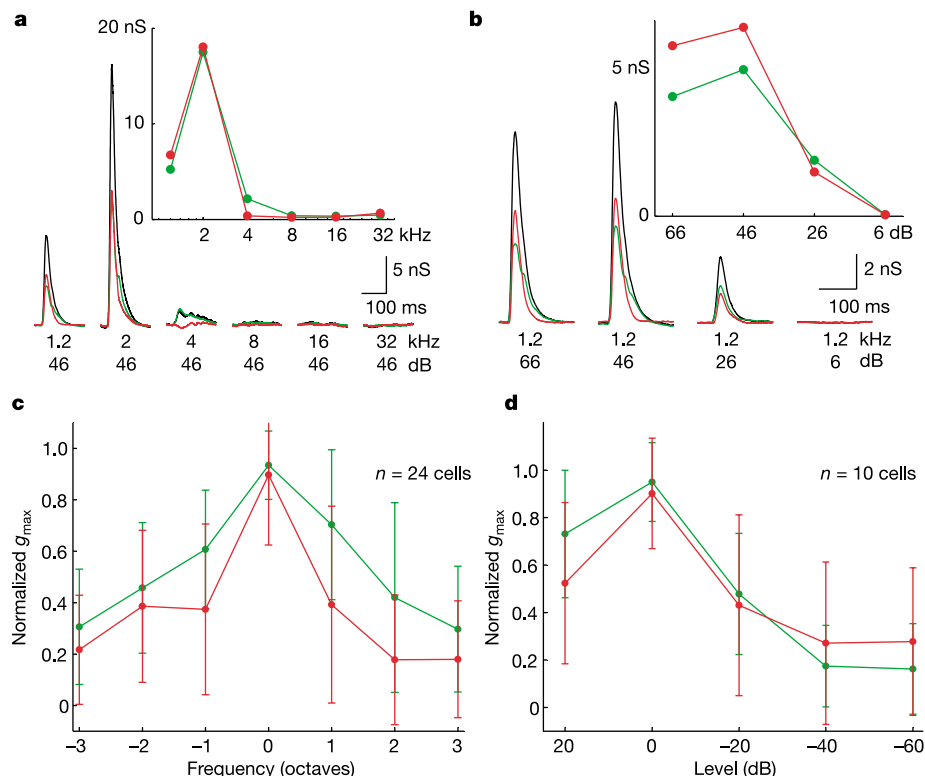


Figure 2 Excitatory, inhibitory, and total synaptic conductances were co-tuned for sound frequency and intensity. **a, b**, The total (black), excitatory (green) and inhibitory (red) synaptic conductances were all tuned for sound frequency (**a**) and intensity (**b**), matching the best frequency (2 kHz) and intensity (46 dB SPL) for this neuron (same cell as in Fig. 1a–c). Insets, Peak excitatory (green) and inhibitory (red) conductances showed similar tuning for frequency (**a**) and intensity (**b**). Error bars are obscured by the filled circles. For this neuron, as in most (29 of 31 cells), the stimulus that elicited maximal depolarization also elicited the maximal excitatory and total synaptic conductance. **c, d**, Co-tuning of normalized mean excitatory and inhibitory conductances for the

population ($n = 229$ stimulus conditions in 31 cells with spikes blocked) against **(c)** ΔF and **(d)** ΔL , where ΔF is the frequency (in octaves) and ΔL is the intensity (in dB) relative to that which evoked maximal total synaptic conductance. For the population, excitation and inhibition were highly correlated (correlation coefficient $r = 0.80$). The ratio of inhibition to excitation tended to be fixed for a given neuron, but varied from one neuron to the next (this ratio was 0.74 ± 0.07 across cells, linear regression slope $\pm 95\%$ confidence, but regression slopes across cells were significantly different from each other; ANCOVA, $P < 0.05$). Of the neurons tested, 70% (7 of 10) were tuned for sound intensity.

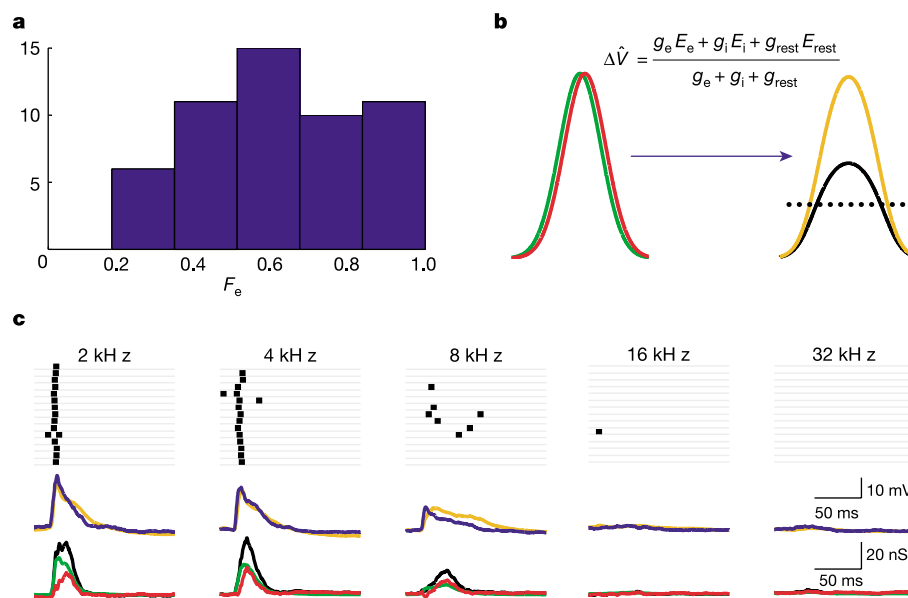


Figure 3 Co-tuned excitation and inhibition governs spiking. **a**, Distribution of the fraction of excitatory conductance $F_e = g_e/(g_e + g_i)$, where g_e and g_i are peak conductances (excitatory and inhibitory, respectively), across all stimuli (responses were included only if the total synaptic conductance was more than 1 nS; mean 0.61 ± 0.21 ($n = 53$ stimulus conditions in 11 cells)). Note that pure inhibition was never observed ($F_e < 0.2$), whereas pure excitation was often observed ($F_e \approx 1$). **b**, Sharpening of tuning can arise from identical excitatory and inhibitory receptive fields. Idealized identical g_e (green) and g_i (red) tuning curves predict a depolarization tuning curve (black) that is sharpened in

comparison with the depolarization produced by excitation alone (orange), especially when a spike initiation threshold (dotted horizontal line) is considered. **c**, For spiking cells, tuning of spikes was co-tuned with, and narrower than, subthreshold responses (same cell and format as in Fig. 1d). In all neurons for which tuning of both spikes and conductances could be assessed, the same stimulus elicited maximal spike count, peak excitatory conductance and peak inhibitory conductance ($n = 16$ cells). For this cell, root-mean-square error (see Supplementary Methods) between the predicted and actual V_m was 4.6 mV (population average 3.0 ± 2.6 ; $n = 21$).

Because inhibition did not seem to be essential to frequency tuning, we wondered whether it was involved in some form of gain control²³, perhaps resulting from the relatively loud (65 dB sound pressure level (SPL)) stimuli that we presented. We therefore tested a wide range of different stimulus intensities (5–65 dB SPL). However, we found that although total synaptic conductances were indeed tuned for sound intensity, the balance of excitation to inhibition remained fixed as intensity varied (Fig. 2b, d). Thus, co-tuning was a general feature of auditory cortical responses to simple tones across a wide range of frequencies and intensities.

Tuning for sound intensity (that is, a non-monotonic response to increasing intensity) has been proposed to result from lateral inhibition²⁴, which would predict that louder sounds should elicit stronger inhibition. Alternatively, intensity tuning might result from an increase in total input conductance¹², in which case louder sounds should elicit a larger total conductance change. However, neither prediction was satisfied; we found non-monotonicity in both excitatory and inhibitory synaptic conductances (Fig. 2b, d) underlying the response for all non-monotonic neurons, indicating that non-monotonicity was inherited from synaptic inputs²⁵. These observations are consistent with the finding that non-monotonic cortical cells are not converted into monotonic cells by inhibitory blockade⁵.

Because inhibition does not seem essential to either frequency or intensity tuning, what is its role? One key function of inhibition might be to sculpt the time course of tone-evoked responses. Excitation and inhibition appeared in a remarkably precise and stereotyped temporal sequence (Fig. 4): inhibition typically followed excitation after a brief delay (Fig. 4b). This sequence is consistent with that seen after electrical stimulation of the medial geniculate body²⁶. In most cells this delay was independent of both frequency and intensity (Fig. 4c, f, g), indicating that relative timing did not usually have a function in frequency or intensity tuning. The delay was stereotyped in the range of 1–4 ms (Fig. 4e; the mean delay

was 2.4 ± 3.6 ms, for $n = 50$ stimulus conditions in 17 cells) even though the latency to the onset of excitation varied, suggesting that inhibition was locked to the excitation rather than to the stimulus itself. Because the period between the onset of excitation and inhibition was short, only a brief window was available for integrating excitatory synaptic inputs. Spikes were elicited only during this brief excitatory window (Figs 1d and 3c). Thus, inhibition acts to enforce the transient^{3,27}, often binary²⁸, nature of stimulus-evoked responses in auditory cortex.

In a minority of cells (2 of 31, or 6%) the delay between excitation and inhibition was tuned for tone frequency (Fig. 4d). In these neurons the maximal depolarization corresponded not to the largest conductance but rather to the longest delay (and the least inhibition). Thus, in a fraction of neurons, timing might have a function in tuning. However, because we did not record spiking responses for these neurons, it is unclear whether these longer delays give rise to sustained firing here or in neurons driven by frequency-modulated sweeps, in which such long and stimulus-dependent delays have also been reported⁹.

Although sensory stimuli elicit both excitation and inhibition in auditory and other sensory systems^{9–14,29}, the consistently short and stereotyped delay we observed between excitation and inhibition, and its relationship to transient spiking, have not previously been reported. In contrast, long and variable delays have been proposed to underlie sweep selectivity in auditory cortex⁹. This difference between short, stereotyped delays and long, variable delays might be due to the differences in stimuli (tones as opposed to sweeps), suggesting that excitatory–inhibitory delays might have different functional roles in shaping the responses to different classes of sounds. Similarly, excitation and inhibition in some simple cells of the primary visual cortex seem to have discrete subregions in space¹¹, yet are co-tuned for orientation¹².

A balance between excitation and inhibition has previously been proposed to account for the apparently random firing of cortical

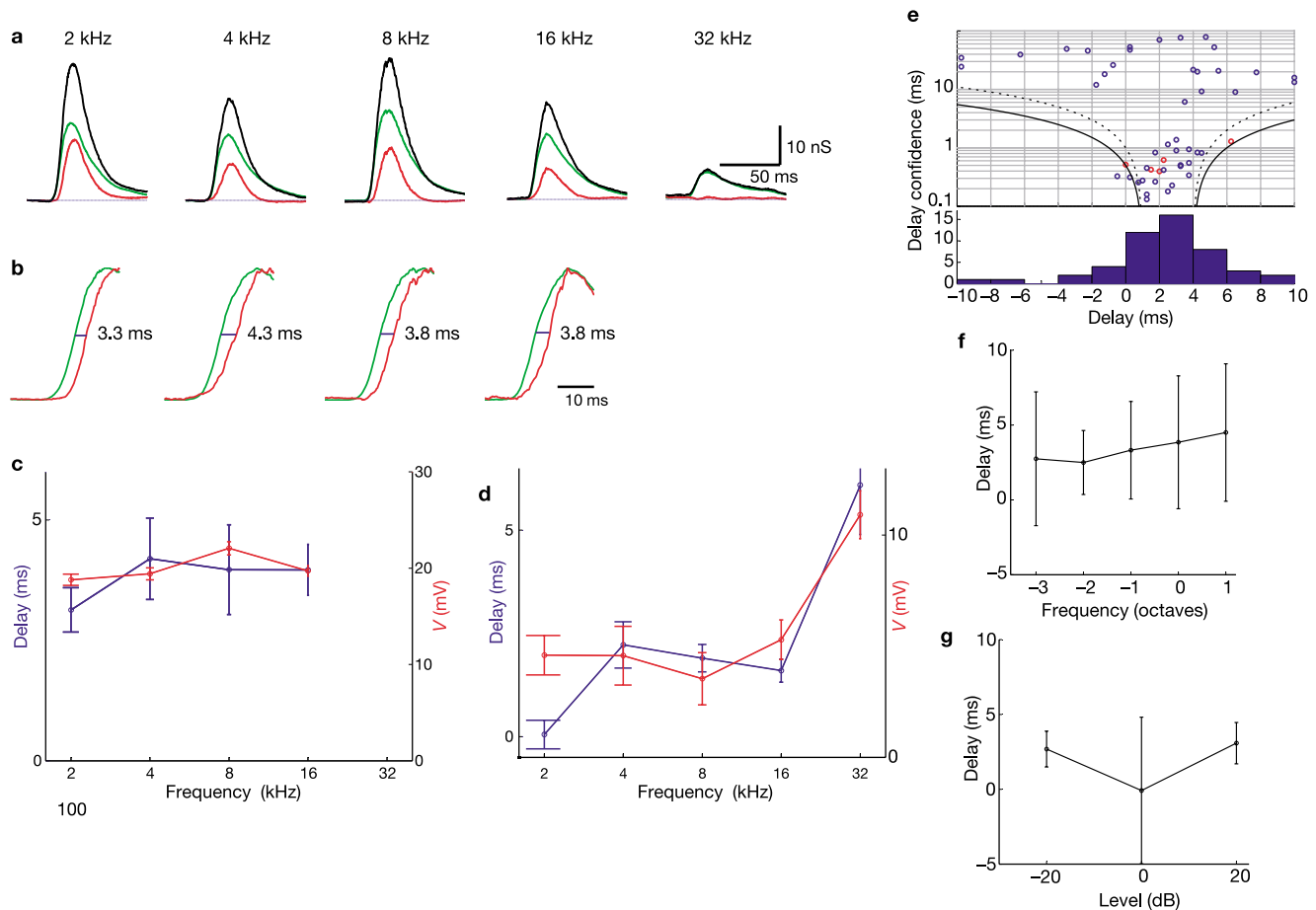


Figure 4 Excitatory conductance preceded inhibitory conductance by a brief delay. **a**, The excitatory conductance (green) led the inhibitory conductance (red) across frequencies (different neuron from those shown in Figs 1–3). **b**, Delays were extracted from normalized conductance waveforms at half maximal amplitude (excitation minus inhibition). **c**, Delays (blue circles, left ordinate) and peak $\Delta\hat{V}_m$ (red circles, right ordinate) for this cell, plotted as a function of frequency. Error bars are standard deviations obtained by bootstrap resampling. **d**, For a different cell, delay followed the same trend as peak $\Delta\hat{V}_m$ as a function of frequency: both were maximal at the best frequency (32 kHz) for this cell (symbols as in **c**). **e**, Bottom: distribution of delays across all cells and stimuli. The mean delay was 2.4 ± 3.6 ms ($n = 17$ cells). For normalization, responses were

included only if both excitation and inhibition were more than 2 nS. Delays of more than 20 ms (corresponding to delays between onset and offset responses) were discarded. Top: for each response, delay is plotted against its confidence level (the standard deviation obtained by bootstrap resampling). Delays with strong confidence (that is, small values on the ordinate) tend to fall in the range 1–4 ms. All but two delays fall within two standard deviations (solid curve) of the range 1–4 ms (dotted line shows one standard deviation), as expected by chance. Responses from the delay-tuned cell in **d** are shown in red. **f**, **g**, Delay plotted against ΔF ($n = 13$ cells) (**f**) and ΔL ($n = 4$ cells) (**g**). Across the population, delay did not show a significant effect of either frequency or intensity (one-way ANOVA).

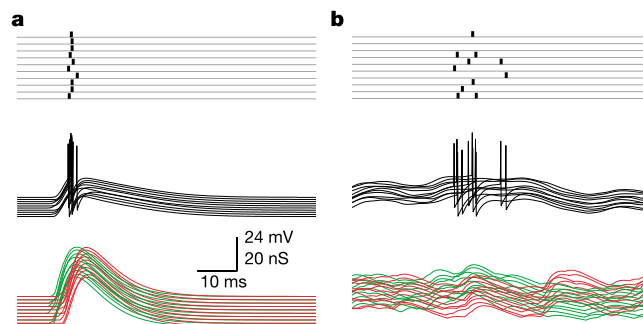


Figure 5 Simulation showing that excitation followed by balanced inhibition increases temporal precision, whereas excitation balanced on average by inhibition decreases temporal precision, using an integrate-and-fire model. **a**, Simulated synaptic input volleys had a mean inhibitory delay of 2.5 ms and a jitter of 1 ms standard deviation. From bottom to top: resulting excitatory (green) and inhibitory (red) synaptic conductances; resulting membrane potential (black lines); and output spike times (black vertical marks) for ten repeated trials. Time jitter in the output spike was 0.6 ms, indicating that inhibitory

quenching increased temporal precision relative to the 1-ms input jitter. **b**, When excitatory and inhibitory synaptic inputs were instead balanced on average over a 100-ms window, again with 1 ms trial-to-trial jitter, firing was highly sensitive to input jitter. Time jitter in the output spike was 4.7 ms, indicating that balanced inhibition increased randomness. (Because inputs in **b** were spread out over a longer time window than in **a**, the number of inputs was increased from 20 to 56 to achieve the same mean spike count of 1.0 per trial.)

neurons by producing large random fluctuations in the membrane potential⁷. In such models, however, it is only the average rate of excitatory and inhibitory postsynaptic potentials that is balanced; the individual events occur at random times, like raindrops. By contrast, the tight coordination we observed reflects a more precise control of cortical circuitry, unlikely to arise from random and independent synaptic activity arriving at a uniform rate³⁰. Moreover, the rapid quenching of excitation by inhibition limits the window available for temporal summation, enabling neurons to behave as coincidence detectors and thereby increasing temporal precision⁶.

We illustrate this distinction by using a simple integrate-and-fire model (Fig. 5). Balanced but delayed inhibition decreased the trial-to-trial jitter in output spike times compared with the jitter in the input spike times (from 1 ms to 0.6 ms; Fig. 5a), whereas excitation and inhibition that were balanced on average instead increased spike time jitter (from 1 ms to 4.7 ms; Fig. 5b). In both of these models, excitation and inhibition are balanced, but in a different sense: a continuing balance produces irregular firing (similar to responses observed in visual cortex, which motivate such models), whereas the transient and temporally offset balance we have observed produces highly transient spiking responses (Figs 1d and 3c), which is consistent with previous findings in auditory cortex^{3,27,28}. Thus our data suggest that balanced inhibition can sharpen neural responses in time, reducing rather than increasing the randomness of cortical operation. □

Methods

We recorded from 62 cells in primary auditory cortex (all subpial depths) of anaesthetized (ketamine-medetomidine) rats aged 17–24 days after birth, using standard blind *in vivo* whole-cell methods. Pure tones had a duration of either 25 ms with a 5 ms 10–90% cosine-squared ramp, or 70 ms with a 20 ms ramp, and were delivered free-field at a rate of 1–2 per second using a calibrated electrostatic speaker in a double-walled sound booth. Total synaptic conductance, corrected for series resistance, was computed assuming an isopotential neuron (see Supplementary Methods for further details).

Received 17 March; accepted 10 October 2003; doi:10.1038/nature02116.

- Shamma, S. A. & Symmes, D. Patterns of inhibition in auditory cortical cells in awake squirrel monkeys. *Hear. Res.* **19**, 1–13 (1985).
- Calford, M. B. & Semple, M. N. Monaural inhibition in cat auditory cortex. *J. Neurophysiol.* **73**, 1876–1891 (1995).
- Sutter, M. L., Schreiner, C. E., McLean, M., O'Connor, K. N. & Loftus, W. C. Organization of inhibitory frequency receptive fields in cat primary auditory cortex. *J. Neurophysiol.* **82**, 2358–2371 (1999).
- Chen, Q. C. & Jen, P. H. Bicuculline application affects discharge patterns, rate-intensity functions, and frequency tuning characteristics of bat auditory cortical neurons. *Hear. Res.* **150**, 161–174 (2000).
- Wang, J., McFadden, S. L., Caspary, D. & Salvi, R. Gamma-aminobutyric acid circuits shape response properties of auditory cortex neurons. *Brain Res.* **944**, 219–231 (2002).
- Pouille, F. & Scanziani, M. Enforcement of temporal fidelity in pyramidal cells by somatic feed-forward inhibition. *Science* **293**, 1159–1163 (2001).
- Shadlen, M. N. & Newsome, W. T. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J. Neurosci.* **18**, 3870–3896 (1998).
- Hartline, H. K., Wagner, H. G. & Ratcliff, F. Inhibition in the eye of *Limulus*. *J. Gen. Physiol.* **39**, 651–673 (1956).
- Zhang, L. I., Tan, A. Y., Schreiner, C. E. & Merzenich, M. M. Topography and synaptic shaping of direction selectivity in primary auditory cortex. *Nature* **424**, 201–205 (2003).
- Borg-Graham, L. J., Monier, C. & Fregnac, Y. Visual input evokes transient and strong shunting inhibition in visual cortical neurons. *Nature* **393**, 369–373 (1998).
- Hirsch, J. A., Alonso, J. M., Reid, R. C. & Martinez, L. M. Synaptic integration in striate cortical simple cells. *J. Neurosci.* **18**, 9517–9528 (1998).
- Anderson, J. S., Carandini, M. & Ferster, D. Orientation tuning of input conductance, excitation, and inhibition in cat primary visual cortex. *J. Neurophysiol.* **84**, 909–926 (2000).
- Monier, C., Chavane, F., Baudot, P., Graham, L. J. & Fregnac, Y. Orientation and direction selectivity of synaptic inputs in visual cortical neurons: a diversity of combinations produces spike tuning. *Neuron* **37**, 663–680 (2003).
- Moore, C. I. & Nelson, S. B. Spatio-temporal subthreshold receptive fields in the vibrissa representation of rat primary somatosensory cortex. *J. Neurophysiol.* **80**, 2882–2892 (1998).
- Carandini, M. & Ferster, D. Membrane potential and firing rate in cat primary visual cortex. *J. Neurosci.* **20**, 470–484 (2000).
- Ojima, H. & Murakami, K. Intracellular characterization of suppressive responses in supragranular pyramidal neurons of cat primary auditory cortex *in vivo*. *Cereb. Cortex* **12**, 1079–1091 (2002).
- Chung, S. & Ferster, D. Strength and orientation tuning of the thalamic input to simple cells revealed by electrically evoked cortical suppression. *Neuron* **20**, 1177–1189 (1998).
- Miller, L. M., Escabi, M. A., Read, H. L. & Schreiner, C. E. Functional convergence of response properties in the auditory thalamocortical system. *Neuron* **32**, 151–160 (2001).
- Schummers, J., Marino, J. & Sur, M. Synaptic integration by V1 neurons depends on location within the orientation map. *Neuron* **36**, 969–978 (2002).

- Phillips, D. P. & Cynader, M. S. Some neural mechanisms in the cat's auditory cortex underlying sensitivity to combined tone and wide-spectrum noise stimuli. *Hear. Res.* **18**, 87–102 (1985).
- Volkov, I. O. & Galazyuk, A. V. Peculiarities of inhibition in cat auditory cortex neurons evoked by tonal stimuli of various durations. *Exp. Brain Res.* **91**, 115–120 (1992).
- Chung, S., Li, X. & Nelson, S. Short-term depression at thalamocortical synapses contributes to rapid adaptation of cortical sensory responses *in vivo*. *Neuron* **34**, 437–446 (2002).
- Chance, F. S., Abbott, L. F. & Reyes, A. D. Gain modulation from background synaptic input. *Neuron* **35**, 773–782 (2002).
- Phillips, D. P. Effect of tone-pulse rise time on rate-level functions of cat auditory cortex neurons: excitatory and inhibitory processes shaping responses to tone onset. *J. Neurophysiol.* **59**, 1524–1539 (1988).
- De Ribaupierre, F., Goldstein, M. H. Jr & Yeni-Komshian, G. Intracellular study of the cat's primary auditory cortex. *Brain Res.* **48**, 185–204 (1972).
- Mitani, A. & Shimokouchi, M. Neuronal connections in the primary auditory cortex: an electrophysiological study in the cat. *J. Comp. Neurol.* **235**, 417–429 (1985).
- Brugge, J. E., Dubrovsky, N. A., Aitkin, L. M. & Anderson, D. J. Sensitivity of single neurons in auditory cortex of cat to binaural tonal stimulation; effects of varying interaural time and intensity. *J. Neurophysiol.* **32**, 1005–1024 (1969).
- DeWeese, M. R., Wehr, M. & Zador, A. M. Binary spiking in auditory cortex. *J. Neurosci.* **23**, 7940–7949 (2003).
- Covey, E., Kauer, J. A. & Casseday, J. H. Whole-cell patch-clamp recording reveals subthreshold sound-evoked postsynaptic currents in the inferior colliculus of awake bats. *J. Neurosci.* **16**, 3009–3018 (1996).
- Stevens, C. F. & Zador, A. M. Input synchrony and the irregular firing of cortical neurons. *Nature Neurosci.* **1**, 210–217 (1998).

Supplementary Information accompanies this paper on www.nature.com/nature.

Acknowledgements We thank K. Miller, L. Miller, M. Kvale, Z. Mainen, M. DeWeese and M. Sutter for comments on an earlier version of this manuscript. This work has been supported by grants to A.M.Z. from the Packard Foundation, the Sloan Foundation, the NIH and the Mathers Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.M.Z. (zador@cshl.edu)

..... An ancient role for nuclear β -catenin in the evolution of axial polarity and germ layer segregation

Athula H. Wikramanayake¹, Melanie Hong^{1,2}, Patricia N. Lee², Kevin Pang², Christine A. Byrum¹, Joanna M. Bince¹, Ronghui Xu¹ & Mark Q. Martindale²

¹Department of Zoology, University of Hawaii at Manoa, 2538 McCarthy Mall, Honolulu 96822, Hawaii

²Kewalo Marine Lab/Pacific Biomedical Research Center, University of Hawaii, 41 Ahui Street, Honolulu 96813, Hawaii

The human oncogene β -catenin is a bifunctional protein with critical roles in both cell adhesion and transcriptional regulation in the Wnt pathway^{1–3}. Wnt/ β -catenin signalling has been implicated in developmental processes as diverse as elaboration of embryonic polarity^{2–6}, formation of germ layers^{4–8}, neural patterning, spindle orientation and gap junction communication², but the ancestral function of β -catenin remains unclear. In many animal embryos, activation of β -catenin signalling occurs in blastomeres that mark the site of gastrulation and endomesoderm formation^{5–10}, raising the possibility that asymmetric activation of β -catenin signalling specified embryonic polarity and segregated germ layers in the common ancestor of bilaterally symmetrical animals. To test whether nuclear translocation of β -catenin is involved in axial identity and/or germ layer formation in 'pre-bilaterians', we examined the *in vivo* distribution, stability and function of β -catenin protein in embryos of the sea anemone *Nematostella vectensis* (Cnidaria, Anthozoa). Here we

show that *N. vectensis* β -catenin is differentially stabilized along the oral–aboral axis, translocated into nuclei in cells at the site of gastrulation and used to specify entoderm, indicating an evolutionarily ancient role for this protein in early pattern formation.

Anthozoans (sea anemones and corals) are basal members of the Cnidaria, a diverse and highly successful phylum, which is the sister group to bilaterian metazoans^{11,12}. Unlike triploblastic bilaterians, which possess derivatives of three embryonic germ layers and are bilaterally symmetrical, cnidarians possess only two germ layers (an external ectodermal layer and internal entodermal layer) and are said to be radially symmetrical around the major longitudinal oral–aboral axis. The evolutionary relationships of the anterior–posterior and dorsoventral axes between bilaterally and radially symmetrical animals remain enigmatic, but as an outgroup to bilaterians, diploblastic cnidarians provide insight into bilaterian development and evolution.

N. vectensis is a small, solitary dioecious sea anemone with a simple life cycle¹³. As with other cnidarian oocytes¹⁴, unfertilized *N. vectensis* oocytes do not appear to have a fixed polarity (that is, animal–vegetal axis). Rather, the adult oral–aboral axis is determined at the time of first cleavage. Gastrulation produces a ciliated, bilayered planula that transforms into an epithelial polyp with an outer ectodermal layer and an inner bifunctional (with both absorptive and contractile properties) entodermal gastrodermis lining the gastric cavity. The molecular mechanisms directing the polarization of an apparently symmetrical cnidarian egg and the segregation of the two germ layers are poorly understood¹⁴, but are critical for understanding the evolution of triploblastic, bilaterally symmetrical animals from a presumed diploblastic, radially symmetrical ancestor¹⁵.

Previous studies indicate a role for the Wnt signalling pathway in axis formation during adult asexual budding in *Hydra*, a derived freshwater hydrozoan¹⁶. However, these studies did not address the role of Wnt signalling during embryogenesis. To determine whether the Wnt signalling pathway functions during early embryogenesis

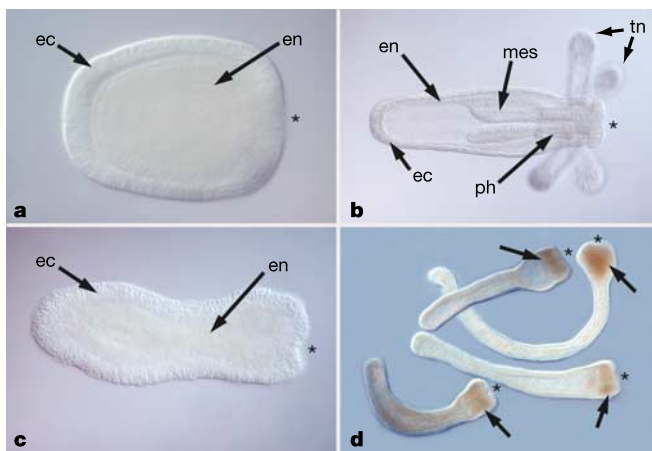


Figure 1 Lithium chloride treatment of *N. vectensis* embryos results in the hyperproliferation of entoderm at gastrulation. **a**, A normal 3-day-old planula stage embryo consisting of an outer ectoderm (ec) and a solid ball of entodermal (en) precursors that have entered through the position of the blastopore (asterisks in all panels). **b**, A 7-day-old juvenile polyp with definitive ectodermal and entodermal germ layers. Entoderm lines the gastric cavity and feeding tentacles (tn) are associated with the site of gastrulation (mouth/anus). The lining of the pharynx (ph) and directive mesenteries (mes) are ectodermal in origin. **c**, A 3-day-old planula treated with 20 mM LiCl from early cleavage stages. The planula elongates as cells continue to enter the blastocoel during gastrulation. **d**, A 7-day-old lithium-treated planula. Both ectodermal and entodermal germ layers have formed, but the ectodermal components of the tentacles and pharynx have failed to form at the expense of entodermal proliferation (arrows).

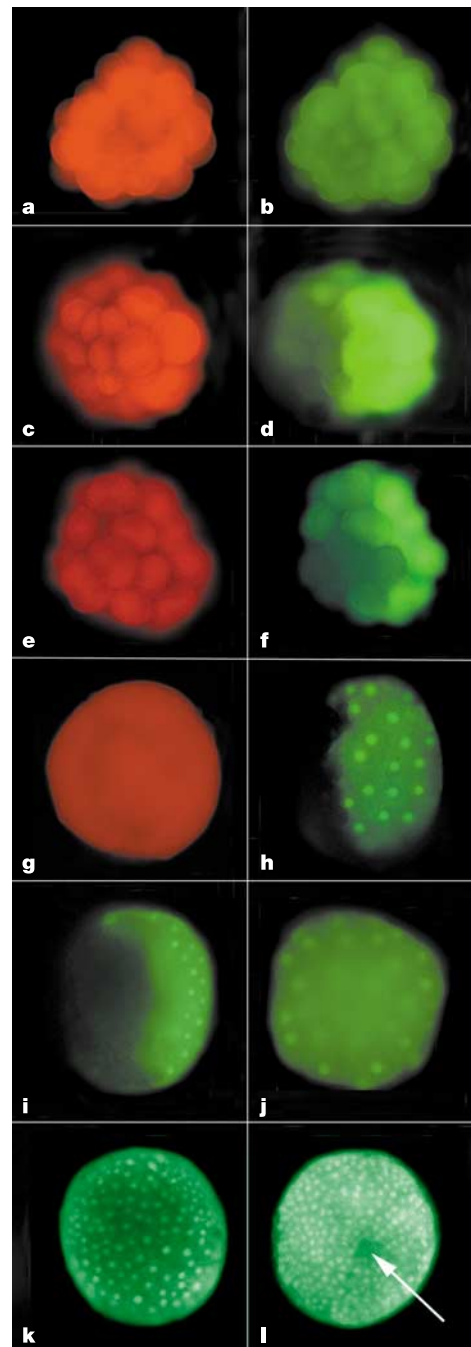


Figure 2 The temporal and spatial dynamics of β -catenin–GFP protein in live *N. vectensis* embryos. mRNA coding for the fusion protein (green) was co-injected with rhodamine dextran (red) into zygotes. All images are oriented with the future oral pole to the right. **a, c, e** and **g** are fluorescent images of the injected embryos seen in **b, d, f** and **h** showing uniform distribution of rhodamine dextran throughout the embryo. **b**, Nv β -cat–GFP protein expression is detected within an hour of injection and is uniformly expressed in the early embryo. At approximately the 32-cell stage, Nv β -cat–GFP protein is degraded at one pole of the embryo, but remains highly expressed at the other pole (**d, f**). GFP fluorescence remains intense at the blastula stage at the pole where gastrulation will occur, and the protein enters the nuclei of entodermal precursors (**h**). **i**, A late blastula stage *N. vectensis* embryo developing from a zygote injected with *Xenopus* β -catenin–GFP chimaeric mRNA. Selective stabilization and nuclear localization of this chimaeric protein is identical to the behaviour of Nv β -cat–GFP. **j**, Injection of a stabilized *Xenopus* β -catenin–GFP mRNA into *N. vectensis* zygotes results in uniform expression and nuclear entry of the protein into all blastomeres of the developing embryos. **k, l**, 20 mM lithium chloride treatment of Nv β -cat–GFP mRNA injected zygotes results in the expanded domains of β -catenin–GFP nuclearization, as expected by GSK-3 β inactivation, at six (**k**) and eight (**l**) hours post injection. The arrow indicates the blastopore in **l**.

in cnidarians, we treated *N. vectensis* zygotes with lithium chloride (LiCl), which interferes with glycogen synthase kinase-3 β (GSK-3 β)-mediated degradation of β -catenin in the Wnt signalling cascade¹⁷. Exposing *N. vectensis* embryos to 10–40 mM LiCl resulted in elongated planulae that fail to make an ectodermal pharynx or tentacles (Fig. 1c, d). Defects in these embryos were first observed during gastrulation with a dramatic increase in endodermal precursors, a classic phenomenon known as “vegetalization”¹⁸.

The effects of lithium on gastrulation movements and germ layer formation led us to follow the spatial and temporal dynamics of β -catenin in live *N. vectensis* embryos using a *N. vectensis* β -catenin (Nv β -catenin):green fluorescent protein fusion protein (Nv β -cat-GFP). Structure-function analysis of the Nv β -catenin molecule has shown that, although many amino- and carboxy-terminal protein-protein interaction domains found in bilaterian β -catenins appear to be absent, both cell adhesion and cell signalling domains are present¹⁹. Microinjection of *in vitro* synthesized Nv β -cat-GFP messenger RNA into fertilized *N. vectensis* eggs results in uniform cytoplasmic expression of this fusion protein throughout the entire embryo during the initial cleavage stages (Fig. 2a, b). However, between the 16-cell and 60-cell stages degradation of Nv β -cat-GFP begins in approximately one-half of the embryo (Fig. 2c–f) as detected by the loss of GFP fluorescence. By the 60-cell stage GFP fluorescence is completely lost in approximately one-half of the embryo, while cells in the other half retain high levels of the fusion protein in the cytoplasm, where it clearly enters the nuclei (Fig. 2g, h). Following GFP expression to later stages showed that Nv β -cat-GFP persists in cells at the future oral pole and remains in cells that enter the blastocoel during gastrulation. Identical results were obtained when chimaeric *Xenopus* β -catenin-GFP mRNA²⁰ was injected into *N. vectensis* embryos (Fig. 2i). Almost identical results have been reported in sea urchin embryos²¹, indicating that a

similar mechanism of protein regulation may have existed in the common ancestor to all eumetazoans.

To verify that the endogenous Nv β -cat protein displays temporal and spatial dynamics similar to those of Nv β -cat-GFP, we used a cross-reactive affinity-purified anti-*Xenopus* β -catenin antibody to localize the protein in early *N. vectensis* embryos (see Supplementary Information). Endogenous β -catenin protein was detected in a subset of nuclei beginning at the 16–32-cell stages (Fig. 3a). By the early and mid-blastula stage, approximately half the blastomeres in the embryos have strong nuclear staining, whereas the remaining blastomeres have little or no nuclear staining (Fig. 3b). The blastomeres with nuclear-localized Nv β -cat at the blastula stage appear to be cells that form entoderm as immunolocalization of Nv β -catenin at the initial stages of gastrulation shows intense staining in invaginating cells (Fig. 3c) as predicted from the dynamics of Nv β -cat-GFP (Fig. 2). Nuclear β -catenin thus provides the first known molecular index of axial asymmetry in *N. vectensis* embryos and this nuclearization is correlated with the site of entoderm formation and gastrulation.

One mechanism of regulating the concentration of β -catenin in cells is the targeting of this protein for degradation by a GSK-3 β - and casein kinase-1-dependent phosphorylation of a cluster of four residues on the amino terminus^{22,23}. Structural analyses have shown that this phosphorylation motif is conserved in β -catenin from humans to cnidarians^{19,23}. To determine whether this aspect of Wnt signalling is conserved in *N. vectensis* we examined the expression and localization of an ‘activated’ form of *Xenopus* β -catenin fused to GFP, in which the GSK-3 β /CK-1 phosphorylation sites had been

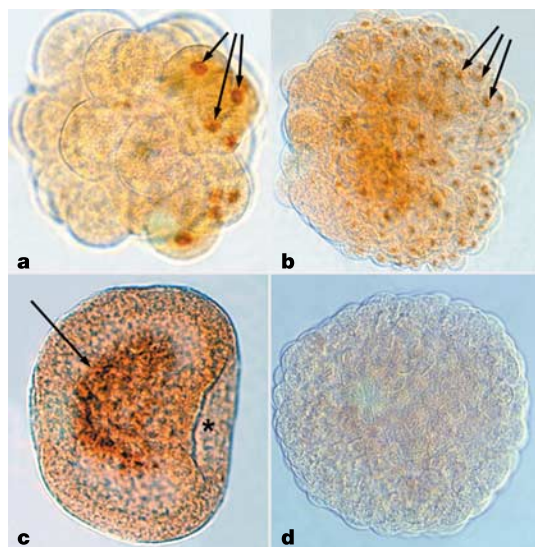


Figure 3 Immunohistochemical localization of endogenous *N. vectensis* β -catenin. **a**, A 32-cell stage *N. vectensis* embryo labelled with an anti- β -catenin antibody. Intense nuclear staining is seen in cells on one side of the embryo but not in the other. **b**, Oblique view of a blastula stage *N. vectensis* embryo labelled with the anti- β -catenin antibody. Nuclear staining is seen on only one side of the embryo corroborating the spatial dynamics of the Nv β -cat-GFP chimaeric protein. **c**, Early gastrula stage embryo labelled with the anti- β -catenin antibody. Intense staining is seen in the developing entoderm. The asterisk marks the blastopore. **d**, A blastula stage embryo that was incubated with the anti- β -catenin antibody following pre-adsorption with a purified bacterially produced Nv β -catenin fusion protein. No staining is seen in the embryos. Arrows point to nuclear β -catenin in **a** and **b** and cytoplasmic β -catenin in the entoderm in **c**.

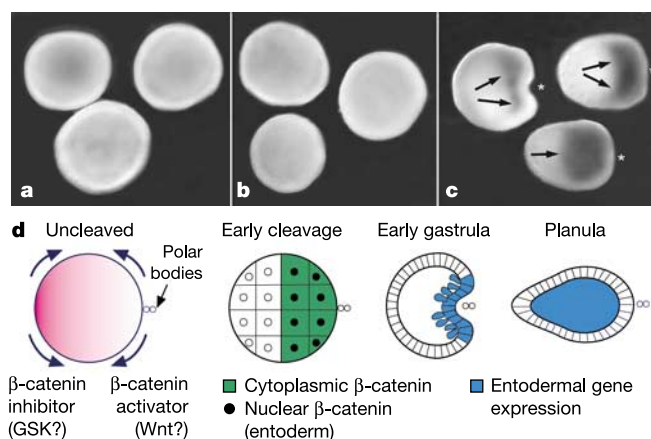


Figure 4 Blocking the nuclear function of β -catenin inhibits entoderm formation in *N. vectensis*. **a**, Injection of *in vitro* synthesized mRNA for sea urchin cadherin—a molecule that binds β -catenin and reduces its free concentration—prevents gastrulation. Injected embryos fail to gastrulate and remain a hollow ball of cells even after 30 h. **b**, Fusion of the transcriptional repression domain from the *Drosophila engrailed* gene to β -catenin generates a chimaeric protein that binds DNA but does not activate downstream targets. The injection of *in vitro* synthesized mRNA coding for this fusion protein also results in the complete absence of gastrulation and entoderm formation. **c**, Control embryos showing entodermal cells (arrows) that have entered the blastocoel through the blastopore (asterisks). The top left embryo is 10 h, the top right embryo 24 h, and the bottom embryo is 30 h post first cleavage. **d**, Model indicating the possible origin of germ layers in a diploblastic embryo. An initial signal generated by the sperm and/or first cleavage polarizes the embryo. This initial polarization event reorganizes the β -catenin destruction machinery, so that it either removes an inhibitor of canonical Wnt signalling from the future oral pole (GSK?) or it localizes an activator of the pathway (Wnt?) to the future oral pole. The stabilization of β -catenin in one hemisphere allows the nuclearization of the protein in cells at the site of gastrulation where it acts as a transcriptional co-activator of specific genes necessary for entoderm specification and gastrulation. The polar bodies mark the future oral pole.

mutated to alanines²². Embryos developing from fertilized eggs injected with an mRNA coding for this chimaeric protein showed GFP-mediated fluorescence in all cells of the blastula, well after it begins to be degraded in normal embryos, and nuclear translocation of Nv β -cat-GFP in all blastomeres (Fig. 2j). Lithium chloride treatment of Nv β -cat-GFP injected zygotes also has the expected result of an expansion of the domain of β -catenin nuclear localization (Fig. 2k, l). These results indicate that much of the complex post-translational regulation of β -catenin destabilization has been conserved in the Metazoa.

In many animal embryos, blocking β -catenin signalling results in failure of germ layer segregation and/or disruption of gastrulation^{2–8,24}. To determine whether β -catenin signalling plays a functional role in early segregation of germ layers in a diploblastic animal, we carried out a loss-of-function experiment designed to competitively titrate β -catenin from the nucleus by overexpressing cadherin^{4–6,24}. Embryos developing from cadherin mRNA-injected zygotes (Fig. 4a) fail to invaginate or gastrulate even after control animals have completed gastrulation (Fig. 4c). To verify that the phenotype was specific to loss of β -catenin signalling, and not a cell adhesion defect caused by cadherin overexpression, we blocked β -catenin-dependent transcription by overexpressing a β -catenin-engrailed repressor-domain fusion protein that has been shown to block β -catenin-dependent transcription²⁵. Injection of β -catenin-engrailed mRNA results in embryos that have phenotypes identical to the cadherin-overexpressing embryos, indicating that the phenotype is generated specifically by loss of β -catenin signalling in *N. vectensis* embryos (Fig. 4b). These data provide compelling evidence that the downstream components of the canonical Wnt signalling pathway are asymmetrically activated during early embryogenesis and are used for the initial segregation of germ layers in a basal cnidarian embryo. We do not know at present whether this asymmetry in nuclear localization of Nv β -catenin is driven by Wnt signals or perhaps by asymmetries created by the fertilizing sperm. It is of some interest that in hydrozoan embryos, sperm can only enter the egg at the site of polar-body formation²⁶, which is the same pole that accumulates nuclear β -catenin in *N. vectensis* and gives rise to endoderm.

Our results indicate a key role for the Wnt/ β -catenin pathway in the evolution of axial asymmetries and endoderm formation in an ancestor that existed before the evolution of the mesodermal germ layer. The lack of a fixed maternal animal–vegetal axis that can serve as a scaffold for axial patterning in cnidarians¹⁴ suggests that the initial molecular asymmetry in these embryos is achieved by a post-fertilization reorganization of the egg cytoplasm that leads to the selective degradation of β -catenin at one pole, and stabilization and subsequent nuclear accumulation in the endoderm-forming pole (Fig. 4d). The observation that similar selective accumulation of β -catenin at one embryonic pole also occurs in deuterostome embryos^{21,27} indicates that this mechanism evolved early during animal evolution and was recruited for generating axial polarities during embryogenesis. We speculate that the evolution of a mechanism to nuclearize β -catenin asymmetrically during early development was the innovation that led to the initial segregation of germ layers from a single-layered blastula-like organism during animal evolution. □

Methods

DNA constructs

A *N. vectensis* PCR fragment corresponding to amino acids 302–656 of mouse β -catenin¹⁹ was digested with *HindIII* and *NotI* to release an 876 base pair (bp) fragment. *N. vectensis* gene-specific primer 1 (5'-TGGGACAGCTTCTGCATCGGCGTGACGGC-3') was used to carry out 5' RACE (rapid amplification of cloned DNA ends) to obtain the 5' end of Nv β -catenin (SMART RACE cDNA amplification kit, Clontech). This fragment was digested with *KpnI* and *HindIII* and used in a three-way ligation with the *HindIII*/*NotI*-digested 3' fragment and *KpnI*/*NotI*-digested pBSKII (Stratagene). This plasmid was used as template for Vent DNA polymerase (NEB)-mediated polymerase chain reaction (PCR) using forward primer 5'-CGGGATCCATGGAGACACGGTATG-3' and reverse primer

5'-CGGGATCCGTTTCGGGAATGTAGTAGG-3' to obtain a Nv β -catenin fragment that contained the translation start site and all the armadillo repeats, but was missing the 3' transactivation domain. The PCR product was digested with *Bam*HI, and ligated into *Bam*HI-digested pCS2 + vector containing the coding sequence for GFP at the *Clal*/*EcoRI* sites to produce pCS2Nv β -cat-GFP. Fidelity of all PCR reactions was verified by sequencing. The activated *Xenopus* β -catenin-GFP construct was made by digesting XBC69 (ref. 22) (a gift from D. Kimelman) with *Clal* and *EcoRI* to release the six Myc tags, and replacing it with GFP. The *Lytechinus variagatus* cadherin⁵, *Xenopus* β -catenin-engrailed²⁵ and *Xenopus* β -catenin-GFP²⁰ constructs have been previously described. All constructs were linearized before mRNA synthesis.

Gamete handling

Fertilized eggs were dejellied with 2% cysteine in one-third strength sea water for 10 min, and rinsed three times in one-third strength sea water in preparation for microinjection or immunohistochemistry.

mRNA synthesis, microinjection and fluorescence microscopy

Linearized constructs were used as templates for SP6-dependent RNA transcription using the mMessage mMachine mRNA synthesis kit from Ambion according to the vendor's specifications. Synthetic mRNA was prepared for microinjection as previously described²⁸. mRNA was injected at a final concentration of 1.0 $\mu\text{g } \mu\text{l}^{-1}$ with 0.2 $\mu\text{g } \mu\text{l}^{-1}$ rhodamine dextran in 40% glycerol under a Zeiss SV-11 fluorescent dissecting microscope. Embryos were raised in one-third strength sea water until the desired stage, at which they were examined under a Zeiss Axioplan compound microscope or a Zeiss 510 laser scanning confocal microscope. Digital images were obtained with either a Zeiss Axiocam or a Nikon Coolpix 990 digital camera.

Immunohistochemistry

Embryos at different stages of development were fixed in paraformaldehyde (MEMPf) and post-fixed with Dent's fixative as described for *Xenopus* embryos²⁹. Briefly, the embryos were initially fixed in MEMPf for 1 h followed by Dent's fixative for 15 min. The embryos were bleached for 30 min in two parts Dent's fixative and one part 30% hydrogen peroxide. Fixed and rinsed embryos were blocked in 5% normal donkey serum/0.05% Tween-20 in phosphate buffered saline (PBS) at pH 7.4. Embryos were incubated in affinity purified anti-*Xenopus* β -catenin antibody³⁰ at 4 °C overnight. Embryos were then rinsed three times in blocking buffer and incubated in horseradish peroxidase-conjugated anti-rabbit donkey secondary antibody (Jackson ImmunoResearch). After 1 h of incubation in the secondary antibody, embryos were rinsed and the staining was developed using DAB substrate. To verify that the primary antibody was recognizing Nv β -catenin, it was pre-adsorbed with a bacterially produced Nv β -catenin fusion protein before incubation with embryos. Rinsed embryos were mounted in methyl salicylate and observed using a Zeiss Axiovert 200 microscope using differential interference contrast optics. Stained embryos were photographed using a Nikon Coolpix 995 digital camera.

Received 25 June; accepted 6 October 2003; doi:10.1038/nature02113.

- Miller, J. R. & Moon, R. T. Signal transduction through beta-catenin and specification of cell fate during embryogenesis. *Genes Dev.* **10**, 2527–2539 (1996).
- Huelsken, J. & Birchmeier, W. New aspects of Wnt signalling pathways in higher vertebrates. *Curr. Opin. Genet. Dev.* **11**, 547–553 (2001).
- Polakis, P. Wnt signalling and cancer. *Genes Dev.* **14**, 1837–1851 (2000).
- Wikramanayake, A. H., Huang, L. & Klein, W. H. beta-catenin is essential for patterning the maternally specified animal-vegetal axis in the sea urchin embryo. *Proc. Natl Acad. Sci. USA* **95**, 9343–9348 (1998).
- Logan, C. Y., Miller, J. R., Ferkowicz, M. J. & McClay, D. R. Nuclear beta-catenin is required to specify vegetal cell fates in the sea urchin embryo. *Development* **126**, 345–357 (1999).
- Imai, K., Takada, N., Satoh, N. & Satou, Y. Beta-catenin mediates the specification of endoderm cells in ascidian embryos. *Development* **127**, 3009–3020 (2000).
- Thorpe, C. H., Schlesinger, A., Carter, J. C. & Bowerman, B. Wnt signalling polarizes an early *C. elegans* blastomere to distinguish endoderm from mesoderm. *Cell* **90**, 695–705 (1997).
- Rocheleau, C. E. et al. Wnt signalling and an APC related gene specify endoderm in early *C. elegans* embryos. *Cell* **90**, 707–716 (1997).
- Schneider, S., Steinbeisser, H., Warga, R. M. & Hausen, P. Beta-catenin translocation into nuclei demarcates the dorsalizing centres in frog and fish embryos. *Mech. Dev.* **57**, 191–198 (1996).
- Miyawaki, K. et al. Nuclear localization of beta-catenin in vegetal pole cells during early embryogenesis of the starfish *Asterina pectinifera*. *Dev. Growth Differ.* **45**, 121–128 (2003).
- Medina, M., Collins, A. G., Silberman, J. D. & Sogin, M. L. Evaluating hypotheses of basal animal phylogeny using complete sequences of large and small subunit rRNA. *Proc. Natl Acad. Sci. USA* **98**, 9707–9712 (2001).
- Collins, A. G. Evaluating multiple alternative hypothesis for the origin of Bilateria: An analysis of 18S rRNA molecular evidence. *Proc. Natl Acad. Sci. USA* **95**, 15458–15463 (1998).
- Hand, C. & Uhlinger, K. R. The culture, sexual and asexual reproduction, and growth of the sea anemone *Nematostella vectensis*. *Biol. Bull.* **182**, 169–176 (1992).
- Goldstein, B. & Freeman, G. Axis specification in animal development. *Bioessays* **19**, 105–116 (1997).
- Knoll, A. & Carroll, S. Early animal evolution: emerging views from comparative biology and geology. *Science* **284**, 2129–2137 (1999).
- Hobmayer, B. et al. WNT signalling molecules act in axis formation in the diploblastic metazoan *Hydra*. *Nature* **407**, 186–189 (2000).
- Klein, P. S. & Melton, D. A. A molecular mechanism for the effect of lithium on development. *Proc. Natl Acad. Sci. USA* **93**, 8455–8459 (1996).
- Kao, K. R. & Elinson, R. P. The legacy of lithium effects on development. *Biol. Cell* **90**, 585–590 (1998).
- Schneider, S. Q., Finnerty, J. R. & Martindale, M. Q. Protein evolution: structure-function relationships of the oncogene beta-catenin in the evolution of multicellular animals. *J. Exp. Zool. B* **295**, 25–44 (2003).

20. Miller, J. R. & Moon, R. T. Analysis of the signalling activities of localization mutants of beta-catenin during axis specification in *Xenopus*. *J. Cell Biol.* **139**, 229–243 (1997).
21. Weitzel, H. E. & Ettensohn, C. A. beta-catenin is differentially degraded along the animal-vegetal axis of early sea urchin embryos in a GSK-3b-dependent manner. *Dev. Biol.* **247**, 479a (2002).
22. Yost, C. *et al.* The axis-inducing activity, stability, and subcellular distribution of beta-catenin is regulated in *Xenopus* embryos by glycogen synthase kinase 3. *Genes Dev.* **10**, 1443–1454 (1996).
23. Ding, Y. & Dale, T. Wnt signal transduction: kinase cogs in a nano-machine? *Trends Biochem. Sci.* **27**, 327–329 (2002).
24. Heasman, J. *et al.* Overexpression of cadherins and underexpression of beta-catenin inhibit dorsal mesoderm induction in early *Xenopus* embryos. *Cell* **79**, 791–803 (1994).
25. Montross, W. T., Ji, H. & McCrea, P. D. A beta-catenin/enrained chimera selectively suppresses Wnt signalling. *J. Cell Sci.* **113**, 1759–1770 (2000).
26. Freeman, G. & Miller, R. L. Hydrozoan eggs can only be fertilized at the site of polar body formation. *Dev. Biol.* **94**, 142–152 (1982).
27. Larabell, C. A. *et al.* Establishment of the dorso-ventral axis in *Xenopus* embryos is presaged by early asymmetries in beta-catenin that are modulated by the Wnt signalling pathway. *J. Cell Biol.* **136**, 1123–1136 (1997).
28. Mao, C.-A. *et al.* Altering cell fates in sea urchin embryos by overexpressing SpOtx, an orthodenticle-related protein. *Development* **122**, 1489–1498 (1996).
29. Klymkowsky, M. W. & Hanken, J. Whole-mount staining of *Xenopus* and other vertebrates. *Methods Cell Biol.* **36**, 419–441 (1991).
30. McCrea, P. D., Briher, W. M. & Gumbiner, B. M. Induction of a secondary body axis in *Xenopus* by antibodies to beta-catenin. *J. Cell Biol.* **123**, 477–484 (1993).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank Y. Marikawa, B. Klein and members of our laboratories for critical reading of the manuscript and for suggestions and P. McCrea for the gift of affinity-purified β -catenin antibody. This work was supported by grants from the NSF and the Hawaii Community Foundation to A.H.W. and by grants from the NSF and NASA to M.Q.M.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.H.W. (athula@hawaii.edu) or M.Q.M. (mqmartin@hawaii.edu).

Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus

Wenhui Li¹, Michael J. Moore¹, Natalya Vasilieva², Jianhua Sui³, Swee Kee Wong¹, Michael A. Berne⁴, Mohan Somasundaran⁵, John L. Sullivan⁵, Katherine Luzuriaga⁵, Thomas C. Greenough⁵, Hyeryun Choe² & Michael Farzan¹

¹Partners AIDS Research Center, Brigham and Women's Hospital, Department of Medicine (Microbiology and Molecular Genetics), ²Perlmutter Laboratory, Pulmonary Division, Children's Hospital, Department of Pediatrics, ³Dana-Farber Cancer Institute, Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA

⁴Tufts University Core Facility, Tufts University School of Medicine, Boston, Massachusetts 02111, USA

⁵Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA

Spike (S) proteins of coronaviruses, including the coronavirus that causes severe acute respiratory syndrome (SARS), associate with cellular receptors to mediate infection of their target cells^{1,2}. Here we identify a metallopeptidase, angiotensin-converting enzyme 2 (ACE2)^{3,4}, isolated from SARS coronavirus (SARS-CoV)-permissive Vero E6 cells, that efficiently binds the S1 domain of the SARS-CoV S protein. We found that a soluble form of ACE2, but not of the related enzyme ACE1, blocked association of the S1 domain with Vero E6 cells. 293T cells transfected with ACE2, but not those transfected with human immunodeficiency virus-1 receptors, formed multinucleated syncytia with cells expressing S protein. Furthermore, SARS-CoV replicated efficiently on ACE2-transfected but not

mock-transfected 293T cells. Finally, anti-ACE2 but not anti-ACE1 antibody blocked viral replication on Vero E6 cells. Together our data indicate that ACE2 is a functional receptor for SARS-CoV.

So far two types of coronavirus surface receptor have been identified⁵. The group II coronavirus mouse hepatitis virus (MHV) uses murine carcinoembryonic antigen-related cell adhesion molecules (CEACAMs), members of the immunoglobulin superfamily of receptors⁶. A number of group I coronaviruses, for example human coronavirus 229E, transmissible gastroenteritis virus and feline infectious peritonitis virus, require the zinc metalloprotease aminopeptidase N (APN, CD13) for entry into their target cells^{7–9}. Recently, a distinct coronavirus has been identified as the aetiological agent of SARS, an acute pulmonary syndrome characterized by an atypical pneumonia that results in progressive respiratory failure and death in close to 10% of cases^{10–13}. Analysis of the SARS-CoV genome suggests that this virus does not belong to any of the three defined coronavirus groups, and that the SARS-CoV S protein is similarly distinct^{14,15}. Similar to the analogous human immunodeficiency virus (HIV) and influenza proteins, the S proteins of some coronaviruses—including MHV and the group III coronavirus infectious bronchitis virus—are cleaved into two subunits (S1 and S2) by a cellular protease in virus-producing cells^{16,17}. The S proteins of other coronaviruses, including those of group I and probably SARS-CoV, are not cleaved in virus-producing cells^{14,15,18}. Nonetheless, S1 and S2 domains of these latter S proteins can be identified through their homology with the S1 and S2 subunits of cleaved coronavirus S proteins. The S1 domains of all characterized coronaviruses mediate an initial high-affinity association with their respective receptors^{19–21}.

The African green monkey kidney cell line Vero E6 permits replication of SARS-CoV¹⁰. We first investigated whether the S1 domain of the SARS-CoV S protein could bind to Vero E6 cells. Figure 1a demonstrates that a protein expressing residues 12–672 of the SARS-CoV S protein, fused to the Fc domain of human immunoglobulin- γ 1 (IgG1; S1-Ig), specifically recognized a moiety present on Vero E6 cells but did not bind human kidney 293T cells. Using this same fusion protein (Fig. 1c) or a carboxy terminally tagged form of the S1 domain (Fig. 1b), a protein band of approximately 110 kDa could be immunoprecipitated from metabolically labelled Vero E6 cells lysed with 0.3% *n*-decyl- β -D-maltopyranoside in phosphate-buffered saline. When the immunoprecipitated protein was incubated with PNGase F, an enzyme that removes *N*-glycosylation, two bands were observed at approximately 80–85 and 100 kDa (Fig. 1c, lanes 3 and 4).

We then analysed the 110 kDa band by trypsin digestion and mass spectrometry. Three human proteins were identified whose sequences were consistent with the masses of tryptic fragments obtained from this band. Two of these proteins, myosin 1b and major vault protein, do not localize to the cell surface and are ubiquitously expressed, therefore we did not analyse them further. Eight independent tryptic fragments consistent with sequences comprising 17% of the amino acid sequence of human ACE2 were also identified (Supplementary Information). Because both the tissue distribution and subcellular localization of ACE2 are appropriate for a receptor for SARS-CoV^{3,22,23}, we cloned it from complementary DNA obtained from human lung for further analysis.

293T cells transfected with plasmid expressing ACE2, but not those transfected with vector alone, were specifically recognized by S1-Ig (Fig. 2a). A soluble form of ACE2, but not that of the related enzyme ACE1, blocked the association of S1-Ig with Vero E6 cells (Fig. 2b). Furthermore, S1-Ig immunoprecipitated a soluble form of ACE2 (Fig. 2c, lane 5) that was recognized by an anti-ACE2 antibody (lane 8). When the approximately 110 kDa soluble form of ACE2 was incubated with PNGase F, an 85 kDa band was observed (Fig. 2c, lanes 5 and 6), indicating that the lower band observed in

PNGase-F-treated Vero E6 immunoprecipitates is ACE2 (Fig. 1c, lanes 3 and 4). These data demonstrate a specific, high-affinity association between the S1 domain of the SARS-CoV S protein and ACE2.

Proteins that mediate fusion between viral and cellular membranes can in some cases also do so between cells that express the viral fusion protein and those that express the viral receptor. For instance, cells expressing the HIV-1 envelope glycoprotein gp160 can form multinucleated syncytia with cells expressing the HIV-1 receptors CD4 and CCR5 (ref. 24). Because syncytia have been observed in Vero E6 cultures infected with SARS-CoV, in SARS-CoV-infected primates, and in SARS patients^{10,12,13}, we investigated whether 293T cells expressing the SARS-CoV S protein could fuse with 293T cells expressing ACE2. Figure 3a shows that these cells formed syncytia efficiently. As expected, 293T cells transfected with CD4 and CCR5 formed many syncytia with cells expressing HIV-1 gp160 (top left), but not with cells expressing SARS-CoV S protein (bottom left). In contrast, 293T cells expressing ACE2 did not form syncytia with cells expressing gp160 (top right), but

formed many large syncytia with cells expressing S protein (bottom right). 293T cells expressing S protein also efficiently formed syncytia with ACE2- but not mock-transfected HeLa cells (not shown). Syncytia formation between S protein- and ACE2-expressing cells was blocked by more than 95% by affinity-purified antibody raised against ACE2, but not by control antibody (Fig. 3b). These data demonstrate that the ability of the SARS-CoV S protein to mediate cell-cell fusion is dependent on the presence of ACE2.

We then investigated the ability of ACE2 to mediate viral replication. ACE2- and mock-transfected 293T cells were incubated in the presence or absence of SARS-CoV. After 2 days in culture, numerous detached, round and floating cells were observed in cultures of infected ACE2-transfected cells (not shown), consistent with viral replication. In contrast, infected mock-transfected cells and uninfected ACE2-transfected cells were indistinguishable from uninfected control cells, and these cells displayed none of the signs of cytopathicity that were observed in infected ACE2-expressing cells. Figure 4 shows that ACE2- but not mock-transfected 293T cells supported efficient replication of SARS-CoV. We measured viral genomic RNA in cell supernatants by semi-quantitative polymerase chain reaction with reverse transcription (RT-PCR). We found that viral genome copies in ACE2-transfected cells increased by more than 100,000-fold in the first 48 h (Fig. 4a, bottom panel).

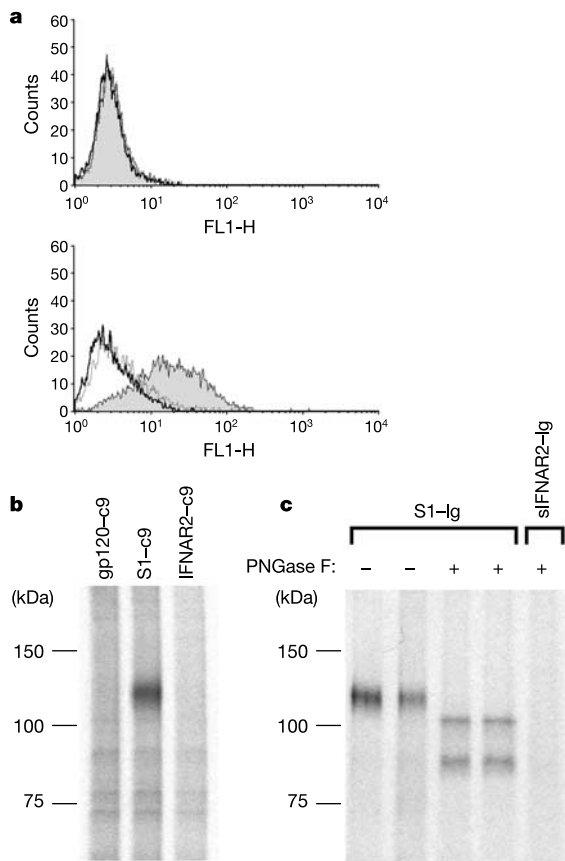


Figure 1 A 110 kDa protein associates with the S1 domain of SARS-CoV S protein. **a**, A fusion protein of the S1 domain (S1-Ig, shaded area), or of the first 327 residues of that domain (S1(327)-Ig, dotted line) with the Fc domain of human IgG1, or culture medium alone (thick line) was incubated with 293T (top panel) or Vero E6 (bottom panel) cells. Binding of fusion proteins to cells was measured by flow cytometry using a FITC-labelled anti-human IgG secondary antibody. **b**, Metabolically labelled Vero E6 cell lysates were immunoprecipitated with HIV-1 gp120, SARS-CoV S1 domain, or the ectodomain of human IFNAR2, each containing a tag (C9) at its C terminus, and an anti-tag antibody. Immunoprecipitates were analysed by SDS-PAGE. **c**, Labelled Vero E6 cell lysates were immunoprecipitated with S1-Ig or soluble IFNAR2-Ig. Immunoprecipitates were treated or not, as indicated, with PNGase F, and analysed by SDS-PAGE.

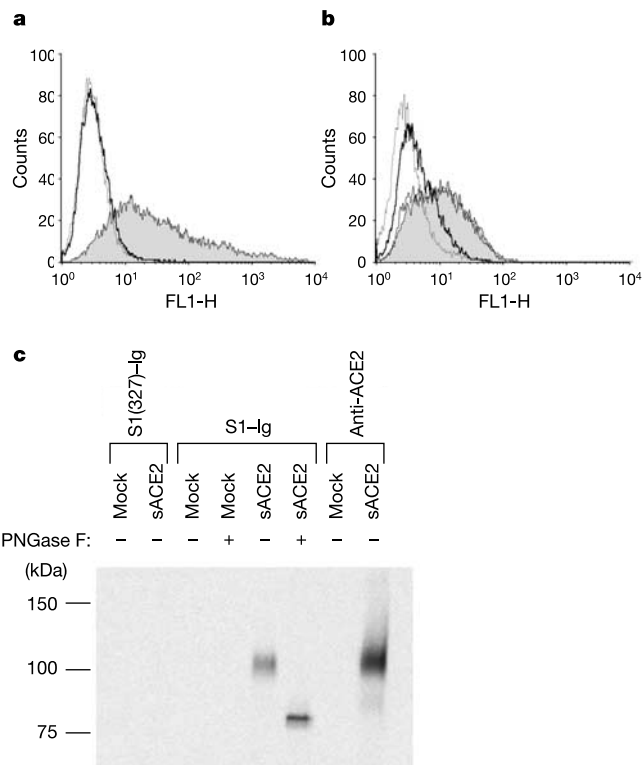


Figure 2 A high-affinity association between ACE2 and the S1 domain. **a**, 293T cells transfected with plasmid encoding ACE2 (shaded area and dotted line), or with vector alone (thick solid line), were analysed by flow cytometry with S1-Ig (shaded area and thick solid line) or S1(327)-Ig (dotted line). **b**, Vero E6 cells were incubated with S1-Ig (shaded area, thin solid line and thick solid line), or with medium alone (dotted line), in the presence of soluble ACE1 (thin solid line) or ACE2 (thick solid line), or without either protein (shaded area and dotted line). **c**, Supernatants of radiolabelled 293T cells transfected with plasmid encoding soluble ACE2 or with vector alone (mock) were immunoprecipitated with S1(327)-Ig, S1-Ig, or an anti-ACE2 antibody. Immunoprecipitates were treated or not, as indicated, with PNGase F, and analysed by SDS-PAGE.

In contrast, genome copies in mock-transfected cells increased tenfold in the same period (top panel), indicating some basal replication on 293T cells. In addition, virus titres in supernatants from infected ACE2- and mock-transfected cells were measured by titration onto Vero E6 cells. As shown in Fig. 4b, virus stock obtained 96 h after infection from ACE2-transfected cells efficiently replicated on Vero E6 cells, with visible cytopathic effect (CPE) up to and including the highest dilution assayed (1:6,561), whereas virus stock obtained from mock-transfected cells induced no CPE beyond a 1:27 dilution. Consistent with the basal replication observed in both infection assays, minor expression of ACE2 on 293T cells was subsequently detected by an anti-ACE2 antibody (Supplementary Information). We also investigated the ability of anti-ACE2 antibody to inhibit viral replication on Vero E6 cells. Anti-ACE1 antibody had no observable effect on virally induced cytopathicity, whereas anti-ACE2 antibody inhibited cytopathicity in a dose-dependent manner (Fig. 4c). Collectively, these data show that ACE2 contributes substantially to the efficiency of SARS-CoV replication.

We show here that ACE2 can be immunoprecipitated from Vero E6 cells by the S1 domain of the SARS-CoV S protein, that it mediates fusion with S-protein-expressing cells, and that it promotes viral replication in a cell line otherwise inefficient for SARS-CoV infection. Moreover, syncytia formation and viral replication were specifically inhibited by an anti-ACE2 antibody. These data demonstrate that ACE2 is a functional receptor for SARS-CoV. Real-time PCR analysis of 72 human tissues has demonstrated efficient expression of *ACE2* messenger RNA in the bronchus and lung parenchyma as well as in the heart, kidney and gastrointestinal tract²³. Lung and kidney are also the primary sites of expression of murine ACE2 (ref. 22). This tissue distribution is consistent with the pathology of SARS, which is characterized by an acute infection of the lungs^{10–12}. SARS-CoV has also been found in kidney tissue of infected patients¹⁰, and it replicates efficiently on the primate kidney cell lines FRhK-4 and Vero E6 (refs 10, 12). Active viral replication in the small and large intestine has been observed²⁵. Detection of virus in the faeces¹¹ may reflect expression of ACE2 message in these tissues²³. The physiological role of ACE2 in most of these tissues has

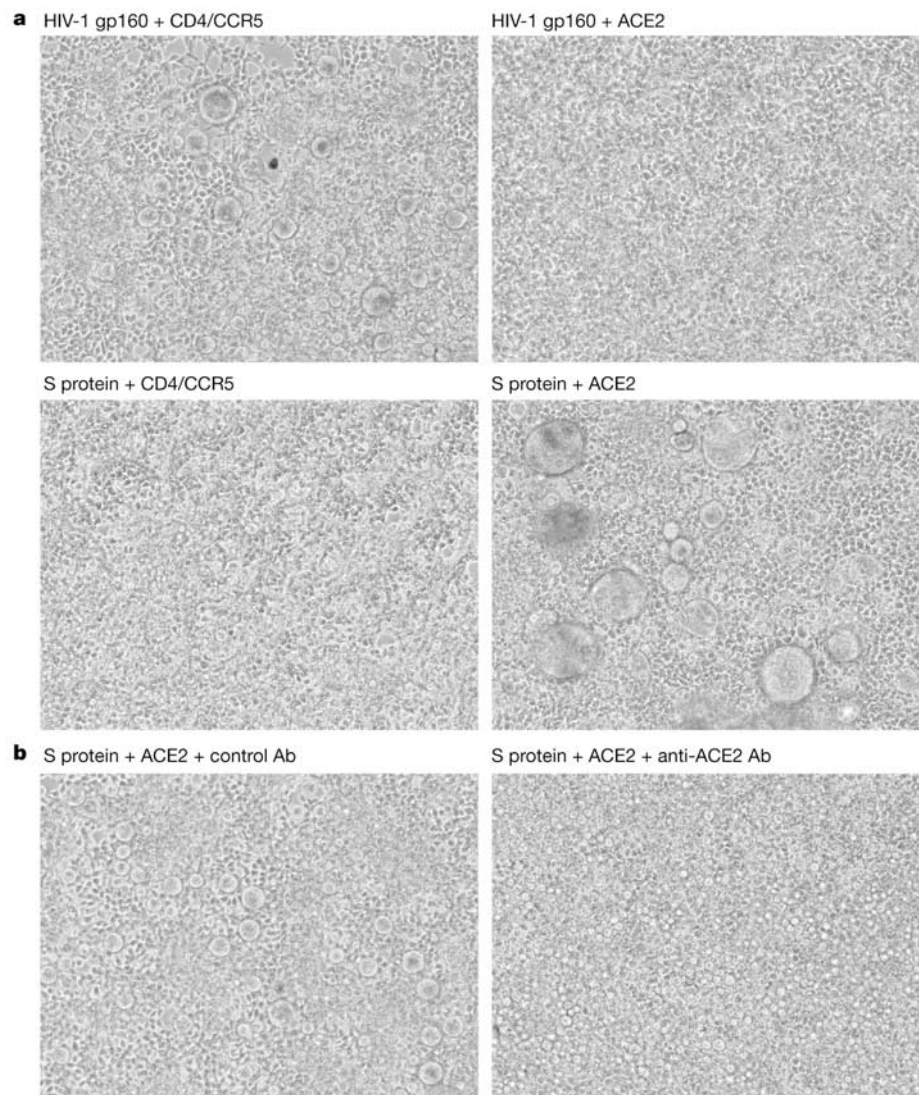


Figure 3 Syncytia formation between S-protein- and ACE2-expressing cells. **a**, 293T cells transfected with plasmids encoding the HIV-1 envelope glycoprotein gp160 (top row) or SARS-CoV S protein (bottom row) were mixed at a 1:1 ratio with 293T cells transfected with plasmids encoding HIV-1 receptors CD4 and CCR5 (left column) or ACE2 (right

column). **b**, 293T cells transfected with ACE2 plasmid were mixed at a 1:1 ratio with 293T cells transfected with plasmid encoding S protein, in the presence of $10 \mu\text{g ml}^{-1}$ goat polyclonal control antibody (left) or goat polyclonal anti-ACE2 antibody (right). Ab, antibody.

not been defined, although ACE2 is thought to be an essential regulator of cardiac function²⁶. Candidate substrates, some of which regulate vasopermeability, have been identified²⁷. It is unclear whether SARS-CoV interferes with ACE2 enzymatic activity in a manner that contributes to the pathogenesis of SARS.

CD13, the receptor for a number of coronaviruses, is, like ACE2, a zinc metalloprotease; however, apart from this commonality the two receptors are widely divergent. This may suggest that some property of this class of proteases contributes to viral replication. However, mutations in the catalytic site of CD13 do not interfere with tissue-culture replication of transmissible gastroenteritis virus,

which uses this enzyme as a receptor^{7,28}. Similarly, our preliminary studies showed that syncytia formation mediated by SARS-CoV S protein remained efficient when two active-site histidines of ACE2 were modified to asparagine (Supplementary Information).

A number of antibodies, peptides and small compounds have been shown to bind to ACE2 (refs 29, 30). It is possible that some of these may be useful in the treatment of SARS, either by blocking the S-protein-binding site, or by inducing a conformation in ACE2 that is unfavourable to binding or fusion. Alternatively, a soluble form of the receptor itself may slow viral replication in an infected individual. Identification of ACE2 as a SARS-CoV receptor will facilitate description of the receptor-binding domain of the S protein, presumably the most effective target epitope of an S1-protein-based subunit vaccine. Also, it is likely that a cell line approved for vaccine production and made permissive for viral replication by ACE2 expression will be the most efficient large-scale producer of a whole-killed or attenuated virus for use as a vaccine. A mouse transgenic for human ACE2 may be useful as an animal model of SARS. Finally, study of the interaction between the SARS-CoV S protein and ACE2 of other animals may provide insights into the origins of the virus. Thus, if SARS returns as a threat to human health, these studies may contribute to its control. □

Methods

Immunoprecipitation and identification of ACE2

Vero E6 cells were metabolically labelled for 48 h with [³⁵S]cysteine and [³⁵S]methionine, and lysed in 1.5 ml per 100-mm dish of 0.3% *n*-decyl- β -D-maltopyranoside (DDM, Anatrace) in phosphate-buffered saline (PBS) containing protease-inhibitor cocktails (Sigma and Roche) and 0.2 mM phenylmethylsulphonyl fluoride (Sigma). After removal of cell debris by centrifugation, lysate was incubated for 1 h at room temperature with 2.5 μ g purified S1-Ig or human interferon- α receptor 2 (IFNAR2)-Ig fusion constructs and protein A Sepharose. Alternatively, C-terminally C9-tagged forms of the S1 domain (S1-C9) or of control proteins (HIV-1 gp120-C9 and IFNAR2-C9) were incubated with the antibody 1D4 (National Cell Culture Center) together with protein A Sepharose. Precipitates were washed twice in 0.3% DDM/PBS and once in PBS alone. Bound proteins were eluted in reducing Laemmli sample buffer at 55 °C or in non-reducing buffer at 37 °C for 10 min. Proteins were separated by SDS-polyacrylamide gel electrophoresis (PAGE) on 8% Tris-glycine gels (Invitrogen). Using this approach approximately 5×10^7 unlabelled Vero E6 cells were used to generate a distinct band of 110 kDa that could be readily visualized by Coomassie staining. This band was excised from the gel and incubated with trypsin, and the masses of tryptic fragments were determined by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Masses were compared with possible tryptic fragments of proteins in the GenBank database using Sequest software.

Receptor expression measurement on Vero E6 and 293T cells

Plasmids encoding the signal sequence of CD5 and a fusion of the S1 domain of the SARS-CoV S protein (residues 12–672), or the first 316 residues of that domain (12–327), with the Fc region of human IgG1 (S1-Ig and S1(327)-Ig, respectively) were transfected into 293T cells, and immunoglobulin fusion proteins were purified on protein A Sepharose beads. A total of 5×10^5 293T cells transfected with ACE2-expressing or control plasmids, or the same number of untransfected Vero E6 cells, were incubated with 15 μ g ml⁻¹ of S1-Ig or S1(327)-Ig in a volume of 100 μ l. In some cases, 15 μ g ml⁻¹ of soluble forms of ACE1 or ACE2 (R&D Systems) was also included. Cells were washed in PBS with 0.5% BSA and 0.1% NaN₃, incubated with FITC-labelled goat anti-human IgG Fc (Sigma), and analysed by flow cytometry.

Syncytia formation

293T cells, approximately 50% confluent on a six-well plate, were transfected by the calcium-phosphate method with plasmids encoding a codon-optimized form of the SARS-CoV S protein or of the HIV-1 envelope glycoprotein gp160 (ADA isolate). In parallel, 293T or HeLa cells were transfected with plasmids encoding ACE2 or the HIV-1 receptors CD4 and CCR5. One day after transfection, cells were trypsinized and those expressing viral-envelope proteins were mixed with cells expressing receptors at a 1:1 ratio, and plated on 12- or 24-well plates. At 24–48 h after mixing, multinucleated giant cells were observed. In some cases 10 μ g ml⁻¹ of goat anti-ACE2 antibody or of goat anti-ACE1 antibody (R&D Systems) was added at the time of cell mixing.

Infection of mock- and ACE2-transfected 293T cells

SARS-CoV (Urbani strain), provided by L. Anderson (Centers for Disease Control and Prevention), was passaged on Vero E6 cells. 293T cells transfected in 25 cm² flasks with pcDNA3.1 alone or pcDNA3.1 expressing ACE2 were infected for 1 h with 1.4×10^3 TCID₅₀ (50% tissue-culture infectious dose) of SARS-CoV, as measured by endpoint titration on Vero E6 cells, or left uninfected, and washed twice in culture medium. Cells were monitored for cytopathic effect for 4 days. Aliquots of cell supernatants were collected 0, 48 and 96 h after wash. RNA was recovered using a viral RNA mini-prep kit

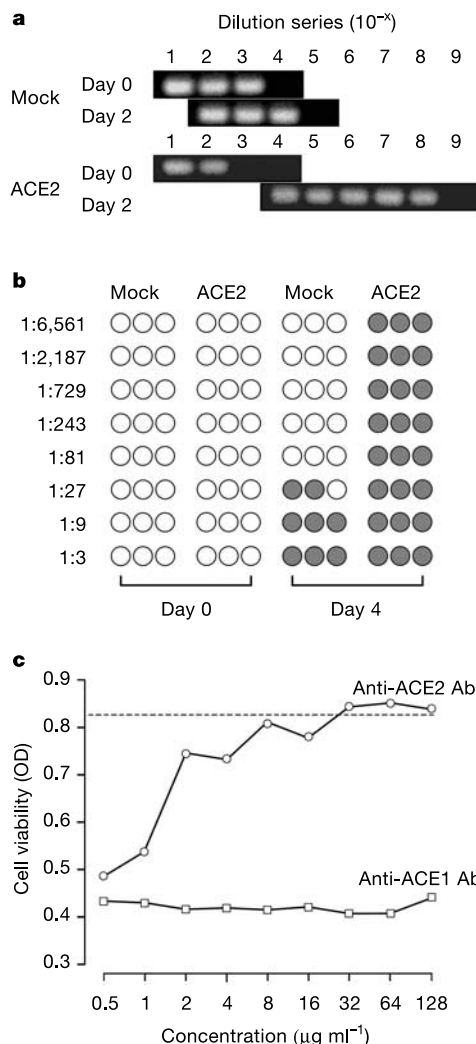


Figure 4 Efficient replication of SARS-CoV in the presence of ACE2. **a**, Mock- or ACE2-transfected 293T cells were infected with SARS-CoV for 1 h, and cell supernatants sampled 0 and 48 h after washing. Viral RNA was measured by RT-PCR at varying dilutions of supernatant. The endpoint of dilution is shown for each group. **b**, Supernatants (collected at days 0 and 4) of infected mock- and ACE2-transfected 293T cells were incubated at the indicated dilutions in triplicate with Vero E6 cells plated on 96-well microtitre plates. Cytopathic effect, indicated by a shaded circle, was monitored for each well 3 days after Vero E6 cell infection. **c**, Duplicate samples of Vero E6 cells were incubated with affinity-purified goat anti-ACE1 or anti-ACE2 antibody, at the indicated concentrations, before infection with SARS-CoV. Cells were washed and cell viability, shown as the average optical density (OD) at 490 nm wavelength, was assayed using CellTiter 96. The dotted line indicates viability of uninfected cells. No toxicity was observed in uninfected cells treated with the maximum concentration of either antibody.

(Qiagen). Semi-quantitative RT-PCR was performed using a nested protocol as described by ref. 11. Positive control RNA was provided by C. Drosten (Bernhard Nocht Institute for Tropical Medicine, National Reference Center for Tropical Diseases). Virus titration was performed by seeding 5×10^5 Vero E6 cells per well in 96-well microtitre plates 1 day before infection. Culture supernatant from infected 293T cells was added to the first set of wells in triplicate and serially diluted. Cells were monitored for CPE 3 days after infection of Vero E6 cells. The effect of affinity-purified goat anti-ACE1 or -ACE2 antibody on SARS-CoV-induced cytopathicity was measured by reading absorbance at 492 nm of cells incubated with CellTiter 96 (Promega).

Received 14 September; accepted 23 October 2003; doi:10.1038/nature02145.

- Gallagher, T. M. & Buchmeier, M. J. Coronavirus spike proteins in viral entry and pathogenesis. *Virology* **279**, 371–374 (2001).
- Holmes, K. V. SARS-associated coronavirus. *N. Engl. J. Med.* **348**, 1948–1951 (2003).
- Donoghue, M. *et al.* A novel angiotensin-converting enzyme-related carboxypeptidase (ACE2) converts angiotensin I to angiotensin 1–9. *Circ. Res.* **87**, E1–E9 (2000).
- Tipnis, S. R. *et al.* A human homolog of angiotensin-converting enzyme. Cloning and functional expression as a captopril-insensitive carboxypeptidase. *J. Biol. Chem.* **275**, 33238–33243 (2000).
- Holmes, K. V. *et al.* Coronavirus receptor specificity. *Adv. Exp. Med. Biol.* **342**, 261–266 (1993).
- Dveksler, G. S. *et al.* Several members of the mouse carcinoembryonic antigen-related glycoprotein family are functional receptors for the coronavirus mouse hepatitis virus-A59. *J. Virol.* **67**, 1–8 (1993).
- Delmas, B. *et al.* Aminopeptidase N is a major receptor for the entero-pathogenic coronavirus TGEV. *Nature* **357**, 417–420 (1992).
- Tresnan, D. B. & Holmes, K. V. Feline aminopeptidase N is a receptor for all group I coronaviruses. *Adv. Exp. Med. Biol.* **440**, 69–75 (1998).
- Yeager, C. L. *et al.* Human aminopeptidase N is a receptor for human coronavirus 229E. *Nature* **357**, 420–422 (1992).
- Ksiazek, T. G. *et al.* A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**, 1953–1966 (2003).
- Drosten, C. *et al.* Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N. Engl. J. Med.* **348**, 1967–1976 (2003).
- Kuiken, T. *et al.* Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome. *Lancet* **362**, 263–270 (2003).
- Fouchier, R. A. *et al.* Aetiology: Koch's postulates fulfilled for SARS virus. *Nature* **423**, 240 (2003).
- Rota, P. A. *et al.* Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* **300**, 1394–1399 (2003).
- Marra, M. A. *et al.* The Genome sequence of the SARS-associated coronavirus. *Science* **300**, 1399–1404 (2003).
- Sturman, L. S. & Holmes, K. V. Proteolytic cleavage of peplomer glycoprotein E2 of MHV yields two 90K subunits and activates cell fusion. *Adv. Exp. Med. Biol.* **173**, 25–35 (1984).
- Jackwood, M. W. *et al.* Spike glycoprotein cleavage recognition site analysis of infectious bronchitis virus. *Avian Dis.* **45**, 366–372 (2001).
- Spaan, W., Cavanagh, D. & Horzinek, M. C. Coronaviruses: structure and genome expression. *J. Gen. Virol.* **69**, 2939–2952 (1988).
- Bonavia, A., Zelus, B. D., Wentworth, D. E., Talbot, P. J. & Holmes, K. V. Identification of a receptor-binding domain of the spike glycoprotein of human coronavirus HCoV-229E. *J. Virol.* **77**, 2530–2538 (2003).
- Breslin, J. J. *et al.* Human coronavirus 229E: receptor binding domain and neutralization by soluble receptor at 37 degrees C. *J. Virol.* **77**, 4435–4438 (2003).
- Kubo, H., Yamada, Y. K. & Taguchi, F. Localization of neutralizing epitopes and the receptor-binding site within the amino-terminal 330 amino acids of the murine coronavirus spike protein. *J. Virol.* **68**, 5403–5410 (1994).
- Komatsu, T. *et al.* Molecular cloning, mRNA expression and chromosomal localization of mouse angiotensin-converting enzyme-related carboxypeptidase (mACE2). *DNA Seq.* **13**, 217–220 (2002).
- Harmer, D., Gilbert, M., Borman, R. & Clark, K. L. Quantitative mRNA expression profiling of ACE 2, a novel homolog of angiotensin converting enzyme. *FEBS Lett.* **532**, 107–110 (2002).
- Choe, H. *et al.* The beta-chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates. *Cell* **85**, 1135–1148 (1996).
- Leung, W. K. *et al.* Enteric involvement of severe acute respiratory syndrome-associated coronavirus infection. *Gastroenterology* **125**, 1011–1017 (2003).
- Crackower, M. A. *et al.* Angiotensin-converting enzyme 2 is an essential regulator of heart function. *Nature* **417**, 822–828 (2002).
- Vickers, C. *et al.* Hydrolysis of biological peptides by human angiotensin-converting enzyme-related carboxypeptidase. *J. Biol. Chem.* **277**, 14838–14843 (2002).
- Delmas, B. *et al.* Determinants essential for the transmissible gastroenteritis virus-receptor interaction reside within a domain of aminopeptidase-N that is distinct from the enzymatic site. *J. Virol.* **68**, 5216–5224 (1994).
- Huang, L. *et al.* Novel peptide inhibitors of angiotensin-converting enzyme 2. *J. Biol. Chem.* **278**, 15532–15540 (2003).
- Dales, N. A. *et al.* Substrate-based design of the first class of angiotensin-converting enzyme-related carboxypeptidase (ACE2) inhibitors. *J. Am. Chem. Soc.* **124**, 11852–11853 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Coderre for assistance with RT-PCR, M. Kirk for editing, and S. H. Wang, E. Kieff, J. Sodroski, C. Gerard and N. Gerard for guidance and helpful conversations.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.C. (hyeryun.choe@tch.harvard.edu) or M.F. (farzan@mcbcr.harvard.edu).

Altered thymic T-cell selection due to a mutation of the ZAP-70 gene causes autoimmune arthritis in mice

Noriko Sakaguchi^{1,2*}, Takeshi Takahashi^{1*}, Hiroshi Hata¹, Takashi Nomura¹, Tomoyuki Tagami¹, Sayuri Yamazaki¹, Toshiko Sakihama¹, Takaji Matsutani³, Izumi Negishi⁴, Syuichi Nakatsuru¹ & Shimon Sakaguchi^{1,2}

¹Department of Experimental Pathology, Institute for Frontier Medical Sciences, Kyoto University, Kyoto 606-8507, Japan

²Laboratory for Immunopathology, RIKEN Research Center for Allergy and Immunology, Yokohama 230-0045, Japan

³Department of Immunology, Shionogi Institute for Medical Science, Settsu, Osaka 566-0022, Japan

⁴Department of Dermatology, Gunma University Hospital, Maebashi, Gunma 371-8511, Japan

* These authors contributed equally to this work

Rheumatoid arthritis (RA), which afflicts about 1% of the world population, is a chronic systemic inflammatory disease of unknown aetiology that primarily affects the synovial membranes of multiple joints^{1–3}. Although CD4⁺ T cells seem to be the prime mediators of RA, it remains unclear how arthritogenic CD4⁺ T cells are generated and activated^{1–3}. Given that highly self-reactive T-cell clones are deleted during normal T-cell development in the thymus, abnormality in T-cell selection has been suspected as one cause of autoimmune disease^{4,5}. Here we show that a spontaneous point mutation of the gene encoding an SH2 domain of ZAP-70, a key signal transduction molecule in T cells⁶, causes chronic autoimmune arthritis in mice that resembles human RA in many aspects. Altered signal transduction from T-cell antigen receptor through the aberrant ZAP-70 changes the thresholds of T cells to thymic selection, leading to the positive selection of otherwise negatively selected autoimmune T cells. Thymic production of arthritogenic T cells due to a genetically determined selection shift of the T-cell repertoire towards high self-reactivity might also be crucial to the development of disease in a subset of patients with RA.

The SKG strain, which spontaneously develops chronic arthritis, is derived from our closed breeding colony of BALB/c mice. Joint swelling with hyperaemia became macroscopically evident in SKG mice at about 2 months of age, initially at a few interphalangeal joints of the forepaws, then progressing in a symmetrical fashion to swelling of other finger joints of the forepaws and hindpaws, and larger joints (wrists and ankles) (Fig. 1a–d). Knee, elbow, shoulder or vertebral joints were rarely affected except for the joint at the base of the tail in aged SKG mice (Fig. 1e). Radiographic examination revealed destruction and fusion of the subchondral bones, joint dislocation, and osteoporosis by 8–12 months of age (Fig. 1f–i). Despite suffering from such severe chronic arthritis, most SKG mice survived well to 1 year of age, generally with more severe arthritides in females (Fig. 1j).

Histology of the swollen joints showed severe synovitis with massive subsynovial infiltration of neutrophils, lymphocytes, macrophages and plasma cells, villus proliferation of synovocytes accompanying pannus formation and neovascularization, and neutrophil-rich exudates in the joint cavity (Fig. 1k, l), with progression of synovocyte proliferation, pannus-eroded adjacent cartilage and subchondral bone. Immunohistochemical staining revealed that CD4⁺ T cells predominantly infiltrated the subsynovial tissue (Fig. 1m, n). As extra-articular manifestations of the disease, most (more than 90%) mice older than 6 months of age developed interstitial pneumonitis with various degrees of perivascular and

peribronchiolar cellular infiltration (Fig. 1o); more than 90% showed infiltration of inflammatory cells in the skin (Supplementary Fig. 1a). Some (10–20%) mice had subcutaneous necrobiotic nodules, not unlike rheumatoid nodules in RA (Fig. 1p), and vasculitides (Supplementary Fig. 1b). SKG mice did not show lymphadenopathy or lupus-like diseases (such as immune-complex glomerulonephritis).

SKG mice developed high titres of rheumatoid factor (RF), autoantibodies specific for type II collagen, antibodies reactive with heat shock protein (HSP)-70 of *Mycobacterium tuberculosis* presumably due to a cross-reaction with a conserved epitope of HSP,

severe hypergammaglobulinaemia and a high concentration of the circulating immune complexes. However, there were no significant titres of anti-DNA antibodies or organ-specific autoantibodies such as those specific for thyroglobulins or gastric parietal cells^{1–3,7} (Fig. 1q–v).

Thus, this spontaneous arthritis in SKG mice resembles RA in clinical and histological characteristics of articular and extra-articular lesions and in serological features^{1–3}.

Transfer of spleen and lymph node T cells, CD4⁺ T cells in particular, from arthritic SKG mice produced similar arthritis in

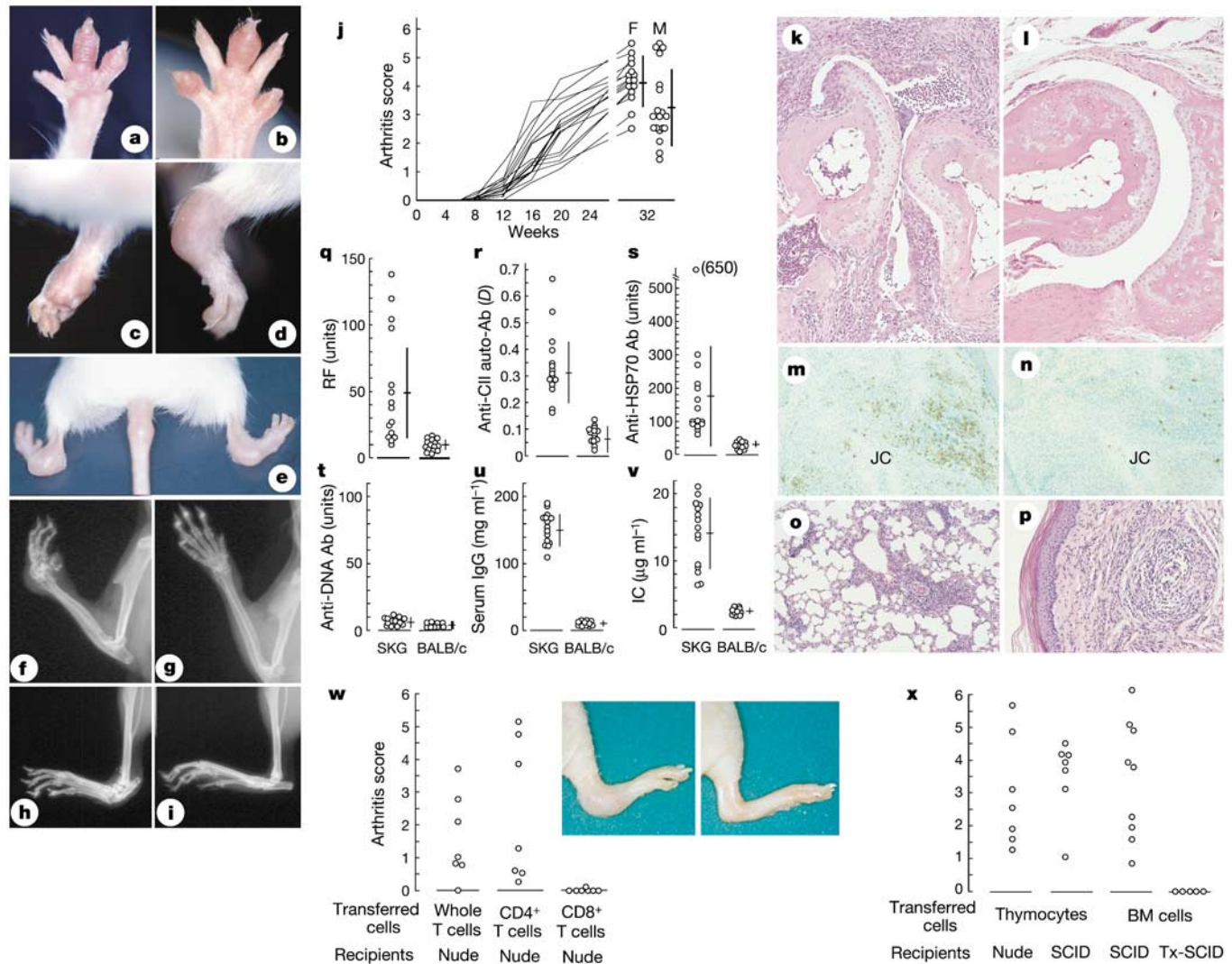


Figure 1 Arthritis in SKG mice. **a–e**, Swelling of joints: fingers (**a**) and toes (**b**) of a 4-month-old SKG mouse; forepaw (**c**) and hindpaw (**d**) of a 6-month-old SKG mouse; deformity of bilateral ankles and swelling of the base of the tail (**e**) in an 18-month-old SKG mouse. **f–i**, X-ray photographs of joints: wrist (**f**) and ankle (**h**) of an 8-month-old SKG mouse, and wrist (**g**) and ankle (**i**) of an 8-month-old BALB/c mouse. **j**, Time course of joint swelling in female SKG mice, and arthritis scores of 8-month-old female (F) and male (M) SKG mice ($n = 15$ each). Vertical bars represent s.e.m. See Methods for details on the scoring of joint swelling. **k–p**, Histology of arthritis and extra-articular lesions in SKG mice: a finger joint of a 6-month-old SKG mouse (**k**) or BALB/c mouse (**l**) (haematoxylin/eosin (HE) staining, original magnification $\times 40$); immunoperoxidase staining of synovial tissue in a finger joint of a 3-month-old SKG mouse with anti-CD4 mAb (GK1.5) (**m**) or anti-CD8 mAb (3–155) (**n**) (magnification $\times 40$) (JC, joint cavity); interstitial pneumonitis (**o**) and a rheumatoid nodule-like lesion (**p**) in an 8-month-old SKG mouse (HE staining, magnification $\times 40$). **q–v**, Serological characteristics of SKG mice: RF (**q**), anti-CII Ab (**r**),

anti-HSP70 antibody (Ab) (**s**), anti-DNA antibody (**t**), concentration of IgG (**u**) and immune complexes (**v**) in the sera from 6-month-old female SKG or BALB/c mice ($n = 15$ each). **w**, Adoptive transfer of arthritis. The same dose (2.5×10^7) of whole, CD4⁺ or CD8⁺ T cells prepared from lymph nodes and spleens of arthritis-bearing 6-month-old SKG mice were transferred to 6-week-old BALB/c nude mice (SLC, Shizuoka). Inset, swelling of a hindpaw (left) and an intact paw (right) of a nude mouse transferred with CD4⁺ T cells and with CD8⁺ T cells, respectively. **x**, Thymocytes (10^6) or T-cell-depleted bone marrow cell suspensions (5×10^6) from the same SKG mice were transferred to 6-week-old nude or C.B.-17 SCID mice (Clea, Tokyo) or SCID mice that had been thymectomized 1 week after birth (Tx-SCID). The severity of arthritis in these mice was assessed 3 months later. CD4⁺ or CD8⁺ T cells were prepared by MACS (Miltenyi Biotec) with more than 95% purity. Bone marrow cells were treated with anti-Thy-1, anti-CD4 and anti-CD8 antibody and rabbit complement before transfer¹⁹.

T-cell-deficient athymic BALB/c nude mice (Fig. 1w), whereas transfer of the sera from the same SKG mice did not (data not shown). Transfer of thymocyte suspensions from arthritic or non-arthritic young SKG mice also elicited severe arthritis in BALB/c nude mice and T/B-cell-deficient C.B-17 severe combined immunodeficiency (SCID) mice (Fig. 1x), whereas thymocyte transfer from normal BALB/c mice did not⁸. Furthermore, transfer of T-cell-depleted bone marrow cell suspensions from SKG mice produced severe arthritis in SCID mice, but not in SCID mice thymectomized before the transfer (Fig. 1x). These results collectively indicate that, first, the RA-like arthritis in SKG mice is a genuine autoimmune disease mediated by apparently joint-specific CD4⁺ T cells; second, the SKG thymus is continuously generating arthritogenic autoimmune T cells; and, third, the SKG bone marrow cells give rise to such arthritogenic T cells through the normal thymic environment, indicating that the arthritogenic abnormality in SKG mice is intrinsic to T cells.

The primary cause of the arthritis in SKG mice is a genetic abnormality, not vertical or horizontal transmission of arthritogenic

microbes. The offspring of crosses between SKG and normal BALB/c mice, whether the mother was SKG or BALB/c, developed no arthritis (Fig. 2a). In contrast, arthritis with a clinical course and severity similar to that in SKG mice occurred in about 50% of the N₂ generation obtained by crossing the non-arthritic F₁ hybrids with SKG. Thus, the genetic abnormality was presumably of a single gene locus (designated the *skg* gene) and inherited in an autosomal recessive fashion with nearly 100% penetrance of the trait in homozygotes raised in our conventional environment.

Linkage analysis between the development of macroscopically evident arthritis and the homozygosity of chromosome-specific microsatellite markers, performed by using the N₂ generation with *Mus musculus castaneus* (CAST/Ei) (that is, SKG × (SKG × CAST/Ei)F₁ mice), mapped the *skg* locus to the centromeric portion of chromosome 1 with the lod score of the locus as infinite (Fig. 2b). Another significant linkage was with the region around the *H-2* locus on chromosome 17 (Fig. 2c). This linkage was observed only when the mice were maintained in a nearly specific-pathogen-free (SPF) condition, but not in arthritis-prone non-SPF conventional

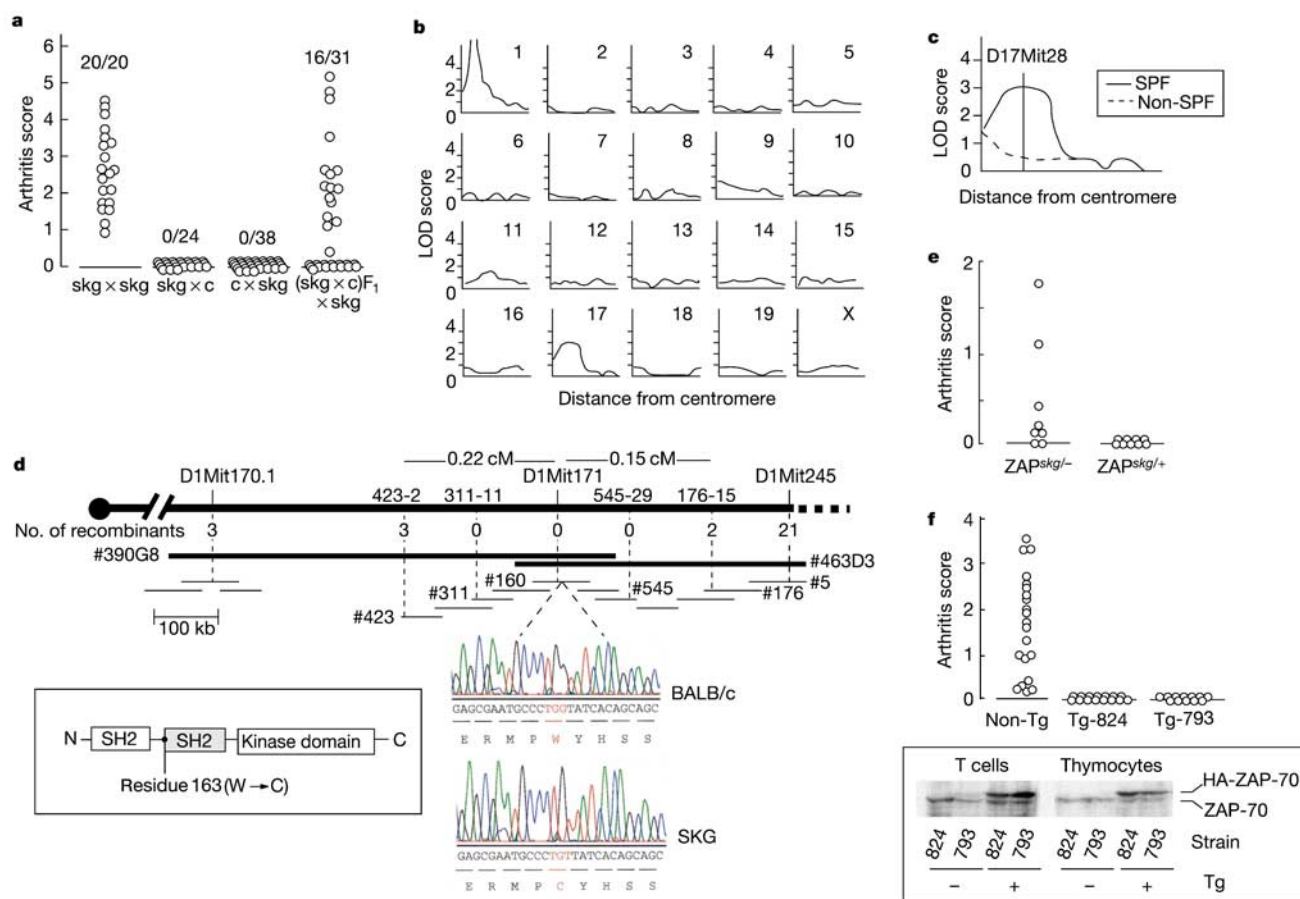


Figure 2 Genetic study of SKG arthritis. **a**, Autosomal recessive inheritance of SKG arthritis. *skg* × *skg*, F₁ of mating SKG females with SKG males; *skg* × *c* and *c* × *skg*, F₁ of mating SKG females with BALB/c males, and BALB/c females with SKG males, respectively; (*skg* × *c*)F₁ × *skg*, N₂ mice of backcrossing (SKG × BALB/c)F₁ mice with SKG mice. **b**, Genome-wide mapping of the *skg* locus. The LOD score curves were obtained by genotyping 58 arthritis-bearing N₂ mice by detecting microsatellite markers (Mouse MapPairs) locating every ~10 cM interval on each chromosome, in accordance with the manufacturer's instruction (Research Genetics). Chromosomes are identified with a number or X in each panel. **c**, Contribution of the loci on chromosome 17 to genetic susceptibility to SKG arthritis in a SPF condition. **d**, Genetic map of the *skg* locus on chromosome 1 with the number of recombinants and physical map of the locus with YAC

and BAC clones. Inset shows the structure of the ZAP-70 protein and the position of the *skg* mutation. The ZAP-70^{W163C} mutation in SKG mice was not found in other mouse strains including C57BL/6, C3H, DBA/1, DBA/2, 129⁺Ter/SV, FvB/N, NOD, MRL-lpr, NZW and CAST/Ei. **e**, SKG mice were mated with heterozygotes of ZAP-70-deficient mice, and the resulting ZAP-70^{*skg*} or ZAP-70^{*skg*+ mice were assessed for joint swelling at 5 months of age. **f**, Rescue of SKG arthritis by expressing human normal ZAP-70 transgene⁹. The arthritis score of Tg-824 or Tg-793 mice, which were backcrossed three times to SKG mice, was assessed at 8 months of age. The expression of transgene-derived human ZAP-70 protein with HA-tag was assessed by western blotting with antibody against human ZAP-70 (Santa Cruz Biotech).}

condition (lod scores at the D17Mit28 locus: 2.95 in SPF compared with 0.40 in non-SPF). In the former condition, homozygosity of the H-2^d haplotype conferred higher genetic susceptibility to the arthritis, indicating that a polymorphism of the major histocompatibility complex (MHC) gene or the gene(s) linked to it can contribute to determining the susceptibility depending on the environmental conditions.

To identify the *skg* gene, we then constructed a high-resolution genetic map of the *skg* region by using 1,352 mice with arthritis among a total of 2,939 N₂ mice (Fig. 2d). The mapping localized the *skg* locus to a 0.37-cM interval between two simple sequence-length polymorphism markers (nos 176-15 and 423-2), covered by two yeast artificial chromosomes (YACs) and ten bacterial artificial chromosomes (BACs) (Fig. 2d, top). Sequencing of several of these BACs localized the ZAP-70 gene on BAC no. 160. Sequencing of the entire coding region of ZAP-70 cDNA from SKG mice and comparison of the sequence with that of BALB/c mice revealed a homozygous G-to-T substitution at nucleotide 489 in the SKG

ZAP-70 gene, which altered codon 163 from tryptophan to cysteine (W163C) (Fig. 2d, bottom). This nucleotide substitution existed in the genomic DNA of SKG mice but not in other strains (see legend to Fig. 2). The position of the mutation corresponds to the initial amino-acid residue of the carboxy-terminal SH2 (SH2C) domain of ZAP-70 (Fig. 2d, inset).

To confirm that the ZAP-70^{W163C} mutation was primarily responsible for SKG arthritis, we crossed SKG mice, which had a ZAP-70^{skg/skg} genotype, with the heterozygotes of ZAP-70-deficient mice (ZAP-70^{+/-}) to produce mice with a ZAP-70^{skg/+} or ZAP-70^{skg/-} genotype⁹ and assessed the development of arthritis in these mice (Fig. 2e). Most of the ZAP-70^{skg/-} mice developed arthritis spontaneously by 3 months of age, in contrast with no arthritis in the ZAP-70^{skg/+} mice. Other immunological abnormalities of T cells and thymocytes were also similar between SKG mice and ZAP-70^{skg/-} mice but these were corrected in ZAP-70^{skg/+} mice (Supplementary Fig. 2a, and see below). Furthermore, transgenic (Tg) expression of the normal human ZAP-70 gene in SKG mice

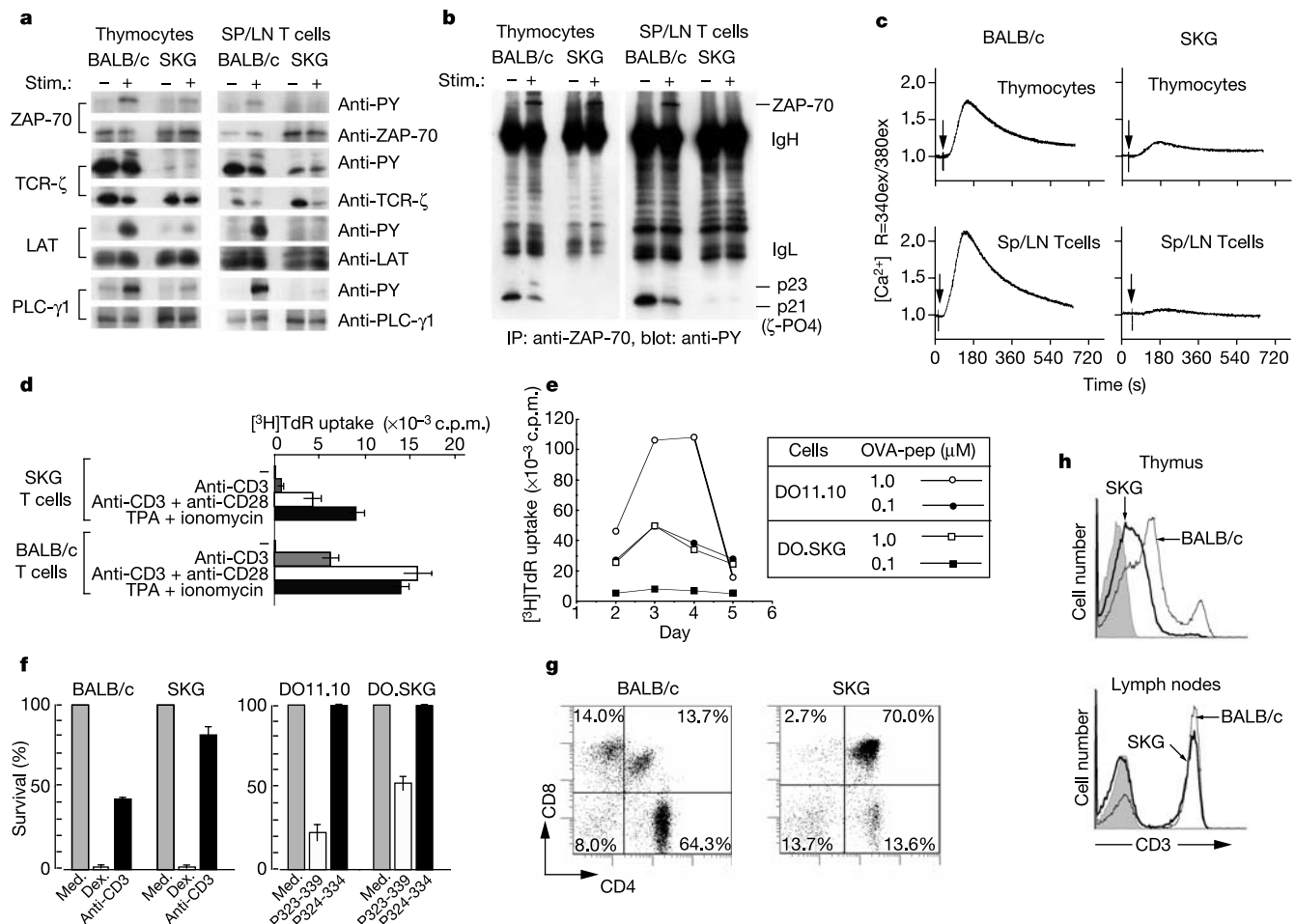


Figure 3 T-cell abnormalities in SKG mice. **a**, Impaired TCR signal transduction in SKG mice. Tyrosine phosphorylation status was assessed by immunoprecipitation and western blotting for ZAP-70, TCR- ζ , LAT and PLC- γ after anti-CD3 mAb stimulation of SKG or BALB/c thymocytes or spleen and lymph node (SP/LN) T cells. **b**, Association between ZAP-70 and TCR- ζ after CD3 crosslinking in SKG T cells and thymocytes. IP, immunoprecipitation. **c**, Ca^{2+} mobilization in SKG T cells or thymocytes after TCR stimulation. **d**, Proliferative responses of SKG T cells to various T-cell stimulations. **e**, Proliferation of CD4⁺ T cells from DO11.10 or DO.SKG mice stimulated with the indicated concentrations of ovalbumin (OVA) peptides in the presence of BALB/c antigen-

presenting cells. **f**, Apoptosis of SKG or BALB/c thymocytes cultured with medium alone (Med.), anti-CD3 mAb (2C11, 1 $\mu\text{g ml}^{-1}$) or dexamethasone (Dex.) (100 nM) (left); apoptosis of DO11.10 or DO.SKG thymocytes cultured with medium alone or indicated ovalbumin peptides (1 μM) (right), as described previously⁹. **g**, Apoptosis of thymocytes in SKG or BALB/c mice injected intraperitoneally with anti-CD3 mAb (2C11; 250 μg) 3 days previously⁹. These figures represent three independent experiments. **h**, Staining of thymocytes or lymph-node cells from a 2-month-old SKG or BALB/c mouse with anti-CD3 mAb.

under the *lck* proximal promoter, which resulted in specific expression of the Tg ZAP-70 protein in thymocytes and peripheral T cells in two Tg lines, completely inhibited the development of arthritis and corrected other immunological abnormalities, contrasting with 100% incidence of arthritis in non-Tg littermates (Fig. 2f and Supplementary Fig. 2b).

Taken together, these results indicate that the ZAP-70^{W163C} alteration is not a mere polymorphism but is the causative mutation of SKG arthritis.

To determine, then, how the ZAP-70 mutation affects T-cell signal transduction, we assessed thymocytes and T cells from SKG or BALB/c mice for the tyrosine-phosphorylation status of major signal transduction molecules, including ZAP-70, TCR- ζ , LAT and PLC- γ 1 (refs 6, 10–12; Fig. 3a, b). After stimulation with anti-CD3 monoclonal antibody (mAb), the levels of phosphorylation of these molecules were extremely low in SKG thymocytes and T cells. Calcium mobilization, one of the earliest events induced by TCR stimulation, was also greatly impaired in SKG thymocytes and T cells, presumably because of the decreased phosphorylation of PLC- γ 1 (Fig. 3c). Immunoprecipitation of ZAP-70 from stimulated SKG thymocytes or T cells failed to co-precipitate tyrosine-phosphorylated p21 and p23 isoforms of TCR- ζ , indicating defective interactions between ZAP-70 and TCR- ζ ^{10,11} (Fig. 3b). Furthermore, downstream mitogen-activated protein (MAP) kinase signalling pathways involving ERK, JNK and p38 MAP kinases were also attenuated in SKG thymocytes¹³ (Supplementary Fig. 3). SKG T cells showed no enhancement of the expression of Syk,

another protein tyrosine kinase that might compensate for the role of ZAP-70 in T cells¹⁴ (T.T., unpublished data). These results collectively indicate that the ZAP-70^{W163C} mutation might impair the recruitment and association of ZAP-70 to tyrosine-phosphorylated immunoreceptor tyrosine-based activation motifs (ITAMs) of TCR- ζ and CD3 chains, thereby altering signal transduction through ZAP-70.

In accord with this attenuated signal transduction through ZAP-70, SKG T cells were hyporesponsive to polyclonal TCR stimulation *in vitro*, for example with anti-CD3 mAb, whereas they responded well to stimulation by 12-*O*-tetradecanoylphorbol-13-acetate (TPA) and ionomycin, which transduces a signal to TCR-distal pathways (Fig. 3d). When the *skg* gene was homozygously introduced to DO11.10 TCR-Tg mice, in which most T cells express Tg TCRs specific for an ovalbumin peptide¹⁵, T cells from such mice (designated DO.SKG mice) required a peptide concentration 10-fold that in DO11.10 T cells to exhibit an equivalent magnitude of proliferation (Fig. 3e). Furthermore, SKG thymocytes were resistant to apoptosis induced by TCR stimulation, but not to dexamethasone-induced apoptosis, as shown by the stimulation of SKG or DO.SKG thymocytes *in vitro* with anti-CD3 mAb or the specific peptide, respectively (Fig. 3f), or by inoculation with anti-CD3 mAb *in vivo* (Fig. 3g).

SKG mice were also aberrant in T-cell development. In comparison with normal BALB/c mice, TCR^{high} mature thymocytes decreased in number and ratio, whereas TCR^{low} immature thymocytes increased in the SKG thymus; the levels of TCR expression on

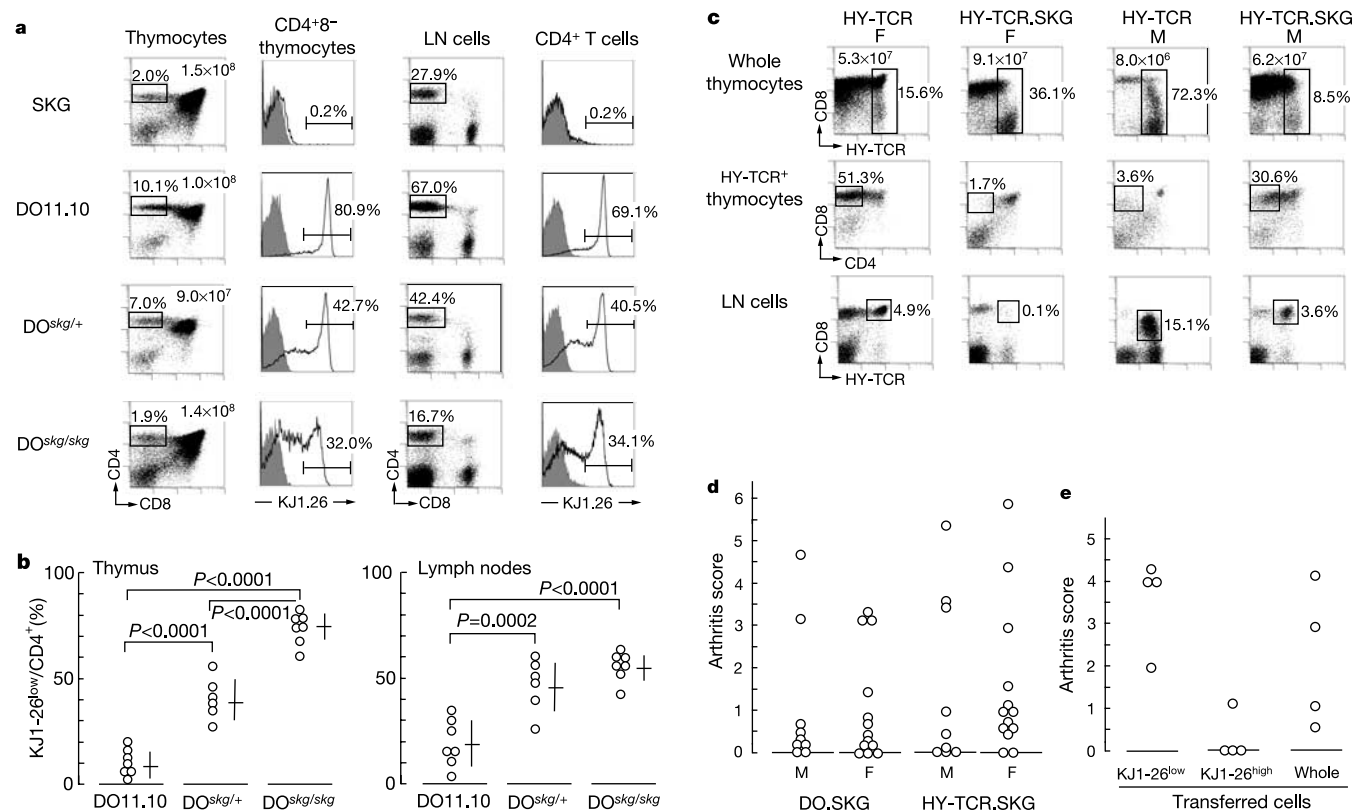


Figure 4 Altered thymic T-cell selection in SKG mice. **a**, Staining of thymocytes or lymph-node cells from a 4-month-old DO11.10, DO^{skg/skg} or DO^{skg/+} mouse with the indicated monoclonal antibodies. Dot-plot figures are scaled logarithmically. Ordinates of histograms denote cell number (arbitrary units). **b**, The percentage of KJ1-26^{low} T cells among CD4⁺CD8[−] thymocytes or CD4⁺ T cells was also assessed for individual DO11.10, DO^{skg/+} or DO^{skg/skg} mice. **c**, Staining of thymocytes or lymph-node cells from a

4-month-old female or male HY-TCR or HY-TCR.SKG transgenic mouse with the indicated monoclonal antibodies. **d**, Arthritis scores of 6-month-old DO.SKG or HY-TCR.SKG mice. **e**, Adoptive transfer of arthritis from arthritis-bearing DO.SKG mice to SCID mice, which were assessed for joint swelling 2 months later. Each fraction of cells (10⁶) was purified by cell sorter from CD4⁺ lymph node and spleen cells.

mature thymocytes or peripheral T cells were comparable between the two strains (Fig. 3h). The numbers of peripheral CD4⁺ and CD8⁺ T cells consequently decreased in SKG mice, with a relative increase in B cells (Supplementary Fig. 4). There was no significant difference in the total number of thymocytes or spleen cells, the thymic architecture, the number and function of natural killer cells (which express ZAP-70) or the composition of T-cell subpopulations expressing particular TCR V α or V β families (Supplementary Fig. 5).

We then assessed the effects of the *skg* mutation on thymic positive and negative selection by introducing the *skg* gene homozygously into TCR-Tg mice^{15,16}. In DO.SKG (that is, DO^{skg/skg}) mice, the number of CD4⁺CD8⁻ mature thymocytes decreased to ~20% of normal DO11.10 mice (Fig. 4a). Furthermore, only ~30% of CD4⁺CD8⁻ thymocytes were Tg-TCR^{high} (when stained with clonotype-specific KJ1-26 mAb)¹⁵ in contrast to ~80% in DO11.10 mice (Fig. 4a, b); the rest of CD4⁺CD8⁻ thymocytes (that is, KJ1-26^{low}, hence Tg-TCR^{low} thymocytes) in DO^{skg/skg} mice seemed to express endogenous TCR α -chains. Similarly, Tg-TCR^{high} T cells constituted only ~30% of CD4⁺ T cells in the periphery of DO^{skg/skg} mice, contrasting with ~70% in DO11.10 mice (Fig. 4a, b); Tg-TCR^{low} T cells in DO^{skg/skg} mice expressed endogenous α -chains associated with transgenic β -chains (Supplementary Fig. 6). In *skg*-heterozygous DO^{skg/+} mice, the decrease or increase in the percentage of Tg-TCR^{high} or Tg-TCR^{low} T cells, respectively, in the thymus and the periphery was intermediate between that in the DO^{skg/skg} and that in DO11.10 mice (Fig. 4a, b).

In HY-TCR Tg mice that express transgenic TCRs specific for male-specific HY antigens on an H-2^b background, female HY-TCR Tg mice positively selected HY-TCR⁺CD8⁺CD4⁻ thymocytes (detected by clonotype-specific T3.70 mAb)¹⁶. In contrast, female HY-TCR Tg mice with the homozygous *skg* gene (designated HY-TCR.SKG Tg mice) on an H-2^b background hardly showed any positive selection of HY-TCR⁺CD8⁺CD4⁻ thymocytes and T cells¹⁷ (Fig. 4c). However, in contrast to substantial deletion of HY-TCR⁺CD8⁺CD4⁻ thymocytes in male HY-TCR Tg mice, male HY-TCR.SKG Tg mice showed efficient positive selection of HY-TCR⁺CD8⁺CD4⁻ thymocytes in almost a comparable number to those in female HY-TCR Tg mice. Furthermore, HY-TCR⁺ peripheral T cells in male HY-TCR.SKG Tg mice expressed nearly normal levels of CD8 expression, contrasting with low-level CD8 expression in male HY-TCR Tg mice¹⁶.

Both DO.SKG and HY-TCR.SKG mice developed arthritis at high incidences (Fig. 4d). Transfer of Tg-TCR^{low} cells (as KJ1-26^{low} cells) from such arthritis-bearing DO.SKG mice to C.B-17 SCID mice induced arthritis, whereas transfer of Tg-TCR^{high} cells did not (Fig. 4e), indicating that TCR-Tg mice with the *skg* mutation positively select arthritogenic T cells expressing endogenous TCR α -chains paired with transgenic β -chains (see above).

Taken together, these results show that the *skg* mutation alters the sensitivity of developing thymocytes to both positive and negative selection, thereby leading to positive selection of otherwise negatively selected autoimmune T cells. This recessive mutation affects T-cell selection even in the heterozygotes, although to a smaller degree than in the homozygotes. Furthermore, the degree of impairment in T-cell signal transduction through ZAP-70 can determine the degree of thymic positive or negative selection, and consequently the phenotype of immunological diseases. The *skg* mutation of the SH2C domain of ZAP-70 elicits autoimmune arthritis, whereas a mutation in the kinase domain of ZAP-70 leads to a severe impairment of positive selection and hence total T-cell deficiency¹⁸.

On the assumption that thymic positive and negative selection of developing T cells requires a certain range of TCR signal through ZAP-70, the *skg* mutation is likely to raise the threshold of TCR avidity for self-peptide/MHC ligands required for the selection to compensate for the reduced signal through the aberrant ZAP-70.

This results in a 'selection shift' of the T-cell repertoire towards high reactivity to self-peptide/MHC ligands in the thymus, leading to the thymic production of highly self-reactive T cells that would not be produced by the normal thymuses of ZAP-70-intact animals. The pathogenic self-reactive T cells thus produced by the SKG thymus apparently overcome the mechanisms of peripheral self-tolerance, for example, mediated by naturally occurring CD25⁺CD4⁺ regulatory T cells^{8,19}. However, it remains to be determined whether the *skg* mutation might also alter the repertoire of CD25⁺CD4⁺ regulatory T cells or somehow affect their suppressive activity^{8,19}.

A critical question on SKG arthritis is why this general alteration in T-cell repertoire should lead to the predominant development of autoimmune arthritis but not other autoimmune diseases. In contrast with other organ-specific autoimmune diseases in which self-reactive T cells destroy the target cells (for example, type 1 diabetes due to destruction of insulin-secreting pancreatic β -cells), a cardinal feature of autoimmune arthritis in SKG mice (and also RA in humans) is that self-reactive T cells do not destroy synovio-cytes but stimulate them to proliferate¹⁻³. This is partly due to a high sensitivity of synovio-cytes to various stimuli that can activate them (as also illustrated by other arthritis models^{7,19-25}) and their distinct capacity to secrete proinflammatory cytokines (such as interleukin-1, interleukin-6 and tumour necrosis factor- α) and chemical mediators that destroy the surrounding cartilage and bone^{1-3,26,27}. It is therefore likely that this unique combination of a high and broad self-reactivity of SKG T cells and a high susceptibility of synovial cells to inflammatory stimuli (including T-cell self-reactivity) leads to predominant development of autoimmune arthritis in SKG mice.

Our findings indicate that mutations of other loci of the ZAP-70 gene or the genes encoding other signalling molecules especially at TCR proximal steps, even if they are heterozygous mutations, might contribute to the development of autoimmune disease by affecting thymic T-cell selection²⁸⁻³⁰. Indeed, we have recently found heterozygous mutations in the ITAM regions of the TCR- ζ chain gene in 2.5% of 160 RA patients in our hospital (H.H., N.S., S.N., T.N. and S.S., unpublished observations). Physical association between these mutated TCR- ζ chain molecules and the normal ZAP-70 molecules, assessed by surface plasmon resonance, was significantly lower than normal as observed between the mutated ZAP-70 in SKG mice and the normal TCR- ζ chain (T.N. and S.S., unpublished data). As RA is a heterogeneous disease with complex genetics¹⁻³, RA with a similar aetiology to SKG arthritis represents only a fraction of the disease. Nevertheless, further analyses of SKG arthritis at each step of the pathogenetic pathway from the ZAP-70 mutation, through thymic generation of autoimmune T cells, to the activation of arthritogenic T cells and inflammatory destruction of the joint, will explain how genetic and environmental factors contribute to the development of RA. This will help in the design of effective methods of detection, treatment and prevention of RA. □

Methods

Scoring of joint swelling

Joint swelling was monitored by inspection and scored as follows: 0, no joint swelling; 0.1, swelling of one finger joint; 0.5, mild swelling of wrist or ankle; 1.0, severe swelling of wrist or ankle. Scores for all fingers of forepaws and hindpaws, wrists and ankles were totalled for each mouse.

Enzyme-linked immunosorbent assay (ELISA)

Affinity-purified mouse IgG (5 μ g ml⁻¹), 10 μ g ml⁻¹ bovine type II collagen (Funakoshi), 5 μ g ml⁻¹ double or single-stranded-DNA, or 10 μ g ml⁻¹ HSP-70 of *Mycobacterium tuberculosis* in PBS, pH 7.2, were used for overnight coating of ELISA plates (Flow Laboratories)^{7,8}. Test sera were diluted 1:10 for anti-type II collagen or anti-HSP-70 assay, 1:20 for RF assay, or 1:40 for anti-DNA assay. Alkaline-phosphatase-conjugated anti-mouse IgG or IgM (for RF assay) (Southern Biotechnology Associates) was used at 1 μ g ml⁻¹ as the secondary reagent^{7,8}. Pooled serum from MRL-lpr mice was used as the standard of arbitrary units in the RF, anti-DNA or anti-HSP70 antibody assay⁷. The titre of anti-CII antibody was expressed as attenuation. The amount of immune complex was measured by anti-C3 assay; the serum concentration of IgG was measured by the single radial immunodiffusion method⁷.

In vitro T-cell activation

SKG or BALB/c T cells (3.0×10^4) were stimulated for 72 h with plate-bound anti-CD3 mAb (2C11) in the presence or absence of plate-bound anti-CD28 mAb in RPMI-1640 medium supplemented with 10% fetal calf serum and 50 μ M 2-mercaptoethanol. Cells were also stimulated with TPA (1.4 ng ml⁻¹) and ionomycin (0.14 μ M). KJ1.26⁺ T cells from DO or DO.SKG mice were stimulated with ovalbumin (323–339) peptide in the presence of X-irradiated syngeneic spleen cells (10^5 cells) as antigen-presenting cells in 96-well round-bottomed plates. The incorporation of [³H]thymidine by proliferating lymphocytes during the last 6 h of the culture was measured.

Immunoprecipitation and western blotting

Thymocytes or purified T cells (5×10^6) were incubated with anti-CD3 mAb (2C11) for 20 min, followed by cross-linking with antibody against Armenian hamster immunoglobulin (Jackson ImmunoResearch) (10 μ g ml⁻¹) at 37 °C for the indicated duration. For immunoprecipitation, cells were lysed with RIPA buffer (0.5% Triton X-100, 20 mM Tris-HCl pH 7.5, 1 mM EDTA, 150 mM NaCl, 20 mM NaF, 1 mM Na₂VO₄) supplemented with protease inhibitors. The immune complexes were recovered by Protein A-conjugated Sepharose beads. For western blot analyses, cells were directly lysed by sample-loading buffer for SDS-polyacrylamide gel electrophoresis (SDS-PAGE) and immediately boiled for 4 min. Recovered immune complexes or total cell lysates were subjected to SDS-PAGE and transferred to poly(vinylidene difluoride) membranes, which were blotted with various antibodies after being blocked with PBS/5% BSA. Antibodies specific for the following proteins were used: ZAP-70 (Santa Cruz Biotech), TCR- ζ (Santa Cruz Biotech), LAT (Upstate Biotechnology), PLC- γ 1 (Upstate Biotechnology), activated or total ERK1/2, SAPK/JNK and p38 MAP kinases (Cell Signalling Technology) or phosphotyrosine (4G10) (Upstate Biotechnology).

Calcium mobilization

Thymocytes or purified T cells were loaded with Fura-2 acetoxymethyl ester (Nacalai) for 30 min at 37 °C. Cells were then incubated with anti-CD3 mAb as described above, washed and resuspended with PBS. CaCl₂ was added to the samples 2 min before cross-linking cell surface-bound anti-CD3 mAb with anti-hamster antibody (Jackson ImmunoResearch). Fura-2 fluorescence was measured by a spectrofluorimeter (Nihon Bunkou).

Received 4 August; accepted 13 October 2003; doi:10.1038/nature02119.

- Harris, E. D. *Rheumatoid Arthritis* (W. B. Saunders, Philadelphia, 1997).
- Feldmann, M., Brennan, F. M. & Maini, R. N. Rheumatoid arthritis. *Cell* **85**, 307–310 (1996).
- Firestein, G. F. in *Textbook of Rheumatology* 5th edn (eds Kelley, W. N., Ruddy, S., Harris, E. D. & Sledge, C. B.) 851–897 (W. B. Saunders, Philadelphia, 1997).
- Marrack, P., Kappler, J. & Kotzin, B. L. Autoimmune disease: why and where it occurs. *Nature Med.* **7**, 899–905 (2001).
- von Boehmer, H. *et al.* Thymic selection revisited: how essential is it? *Immunol. Rev.* **191**, 62–78 (2003).
- Chan, A. C., Iwashima, M., Turck, C. W. & Weiss, A. ZAP-70: a 70 kd protein-tyrosine kinase that associates with the TCR zeta chain. *Cell* **71**, 649–662 (1992).
- Sakaguchi, S. & Sakaguchi, N. Thymus and autoimmunity: capacity of the normal thymus to produce pathogenic self-reactive T cells and conditions required for their induction of autoimmune disease. *J. Exp. Med.* **172**, 537–545 (1990).
- Itoh, M. *et al.* Thymus and autoimmunity: production of CD25⁺ CD4⁺ naturally anergic and suppressive T cells as a key function of the thymus in maintaining immunologic self-tolerance. *J. Immunol.* **162**, 5317–5326 (1999).
- Negishi, I. *et al.* Essential role for ZAP-70 in both positive and negative selection of thymocytes. *Nature* **376**, 435–438 (1995).
- van Oers, N. S. *et al.* The 21- and 23-kD forms of TCR zeta are generated by specific ITAM phosphorylations. *Nature Immunol.* **1**, 322–328 (2000).
- Iwashima, M. *et al.* Sequential interactions of the TCR with two distinct cytoplasmic tyrosine kinases. *Science* **263**, 1136–1139 (1994).
- Zhang, W., Sloan-Lancaster, J., Kitchen, J., Triple, R. P. & Samelson, L. E. LAT: the ZAP-70 tyrosine kinase substrate that links T cell receptor to cellular activation. *Cell* **92**, 83–92 (1998).
- Rincon, M. MAP-kinase signaling pathway in T cells. *Curr. Opin. Immunol.* **13**, 339–345 (2001).
- Noraz, N. *et al.* Alternative antigen receptor (TCR) signaling in T cells derived from ZAP-70-deficient patients expressing high levels of Syk. *J. Biol. Chem.* **275**, 15832–15838 (2000).
- Murphy, K. M., Heimberger, A. B. & Loh, D. Y. Induction by antigen of intrathymic apoptosis of CD4⁺CD8⁺TCR⁰ thymocytes *in vivo*. *Science* **250**, 1720–1723 (1990).
- Kisielow, P., Bluthmann, H., Staerz, U. D., Steinmetz, M. & von Boehmer, H. Tolerance in T-cell-receptor transgenic mice involves deletion of nonmature CD4⁺8⁺ thymocytes. *Nature* **333**, 742–746 (1988).
- Kisielow, P., Teh, H. S., Bluthmann, H. & von Boehmer, H. Positive selection of antigen-specific T cells in thymus by restricting MHC molecules. *Nature* **335**, 730–733 (1988).
- Wiest, D. L. *et al.* A spontaneously arising mutation in the DLAARN motif of murine ZAP-70 abrogates kinase activity and arrests thymocyte development. *Immunity* **6**, 663–671 (1997).
- Sakaguchi, S., Sakaguchi, N., Asano, M., Itoh, M. & Toda, M. Immunologic self-tolerance maintained by activated T cells expressing IL-2 receptor α -chains (CD25). Breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. *J. Immunol.* **155**, 1151–1164 (1995).
- Keffer, J. *et al.* Transgenic mice expressing human tumor necrosis factor: a predictive genetic model of arthritis. *EMBO J.* **10**, 4025–4031 (1991).
- Koussoff, V. *et al.* Organ-specific disease provoked by systemic autoimmunity. *Cell* **87**, 811–822 (1996).
- Pals, S. T., Radaszkiewicz, T., Roozendaal, L. & Gleichman, E. Chronic progressive polyarthritis and other symptoms of collagen vascular disease induced by graft-versus-host reaction. *J. Immunol.* **134**, 1475–1482 (1985).
- Nishimura, H., Nose, M., Hiai, H., Minato, N. & Honjo, T. Development of lupus-like autoimmune

diseases by disruption of the PD-1 gene encoding an ITIM motif-carrying immunoreceptor.

- Immunity* **11**, 141–151 (1999).
- Horai, R. *et al.* Development of chronic inflammatory arthropathy resembling rheumatoid arthritis in interleukin 1 receptor antagonist-deficient mice. *J. Exp. Med.* **191**, 313–320 (2000).
 - Atsumi, T. *et al.* A point mutation of Tyr-759 in interleukin 6 family cytokine receptor subunit gp130 causes autoimmune arthritis. *J. Exp. Med.* **196**, 979–990 (2002).
 - Dayer, J. M. & Burger, D. Cytokines and direct cell contact in synovitis: relevance to therapeutic intervention. *Arthritis Res.* **1**, 17–20 (1999).
 - Naka, T., Nishimoto, N. & Kishimoto, T. The paradigm of IL-6: from basic science to medicine. *Arthritis Res.* **4** (Suppl. 3), S233–S242 (2002).
 - Werlen, G., Hausmann, B., Naecher, D. & Palmer, E. Signaling life and death in the thymus: timing is everything. *Science* **299**, 1859–1863 (2003).
 - Nambiar, M. P. *et al.* T cell signalling abnormalities in systemic lupus erythematosus are associated with increased mutations/polymorphisms and splice variants of T cell receptor zeta chain messenger RNA. *Arthritis Rheum.* **44**, 1336–1350 (2001).
 - Takeuchi, T. *et al.* TCR zeta chain lacking exon 7 in two patients with systemic lupus erythematosus. *Int. Immunol.* **10**, 911–921 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank H. von Boehmer for transgenic mice; M. Singh (GBF, Germany) for a gift of recombinant HSP-70 through the support of the UNDP/World Bank/WHO Special Program for Research and Training on Tropical Diseases; F. Melchers, R. Zinkernagel, K. Yamamoto, R. Suzuki and Z. Fehervari for discussion; A. Kosugi and T. Nakayama for reagents; and E. Morizumi for immunohistochemistry and histology. This work was supported by grants-in-aid from the Ministry of Education, Science, Sports and Culture, the Ministry of Human Welfare, Japan Science and Technology Corporation, and the Organization for Pharmaceutical Safety and Research of Japan.

Authors' contributions The SKG strain was established by S.S. and N.S. The experiments in Figs 1, 2a–d, 3f–h, 4, Supplementary Figs 1, 4 and 6 were conducted by N.S. and S.S.; those in Fig. 2b–d by Ta.T., N.S., H.H., To.T., S.Y., T. S., S.N. and S.S.; those in Figs 2e, f, 3a–c, e and Supplementary Fig. 2 by Ta.T.; that in Fig. 3d by To.T.; that in Supplementary Fig. 3 by T.N., and that in Supplementary Fig. 5 by T.M. ZAP-70-deficient mice were provided by I.N.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.S. (shimon@frontier.kyoto-u.ac.jp).

A positive-feedback-based bistable 'memory module' that governs a cell fate decision

Wen Xiong & James E. Ferrell Jr

Department of Molecular Pharmacology, Stanford University School of Medicine, Stanford, California 94305-5174, USA

The maturation of *Xenopus* oocytes can be thought of as a process of cell fate induction, with the immature oocyte representing the default fate and the mature oocyte representing the induced fate^{1,2}. Crucial mediators of *Xenopus* oocyte maturation, including the p42 mitogen-activated protein kinase (MAPK) and the cell-division cycle protein kinase Cdc2, are known to be organized into positive feedback loops³. In principle, such positive feedback loops could produce an actively maintained 'memory' of a transient inductive stimulus and could explain the irreversibility of maturation^{3–6}. Here we show that the p42 MAPK and Cdc2 system normally generates an irreversible biochemical response from a transient stimulus, but the response becomes transient when positive feedback is blocked. Our results explain how a group of intrinsically reversible signal transducers can generate an irreversible response at a systems level, and show how a cell fate can be maintained by a self-sustaining pattern of protein kinase activation.

Immature *Xenopus* oocytes are arrested in a G2-like phase of the cell cycle. In response to steroid hormones, the oocyte is released

from its G2 arrest, undergoes germinal vesicle breakdown (GVBD), completes meiosis I, proceeds directly into meiosis II, and then arrests again in the metaphase of meiosis II. Early work indicated that at some point in this process, the oocyte irrevocably commits to maturation⁷⁻⁹. We corroborated these observations by using low concentrations of the steroid hormone progesterone and thorough washing procedures to minimize the possibility that commitment was an artefact of inadequate washing. We found that oocytes generally committed to maturation after 2–4 h of progesterone treatment and before (and in one experiment just before) GVBD (Fig. 1a–c). This timing agrees well with studies of the acquisition of ‘maturation inertia’ in *Rana temporaria* oocytes treated with gonadotropin⁷. Oocytes remained arrested in their mature, GVBD state for up to several days after progesterone removal. Thus, transient treatment with progesterone results in a sustained cellular response.

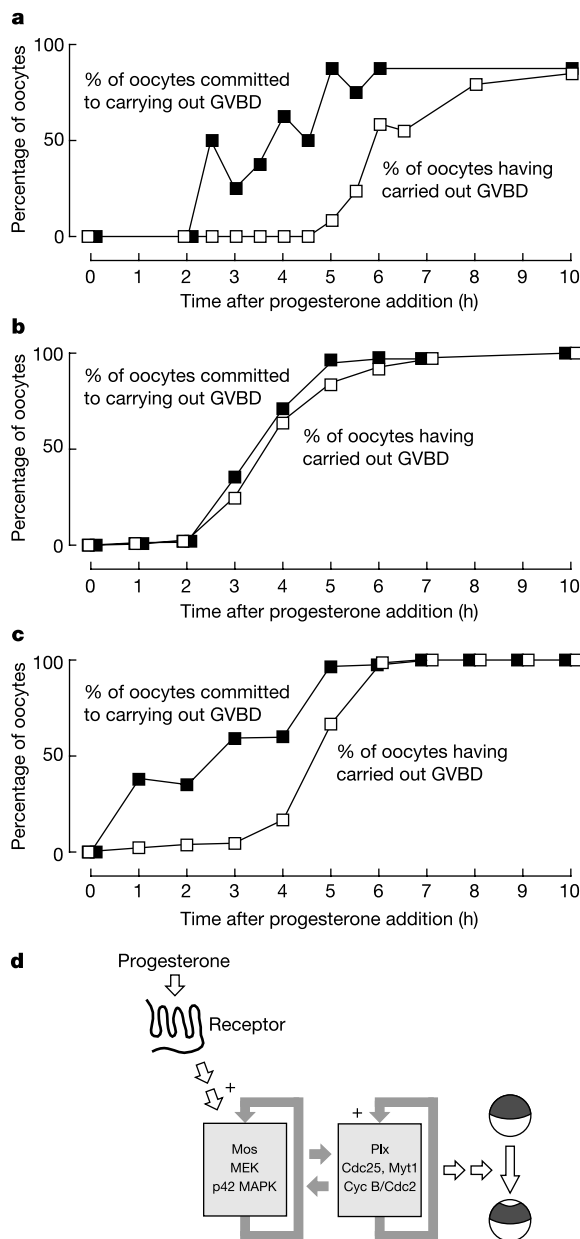


Figure 1 Cell fate commitment during oocyte maturation. **a–c**, Timing of commitment versus maturation from three independent experiments. **d**, View of the signal transduction network that culminates in *Xenopus* oocyte maturation.

At a biochemical level, oocyte maturation is controlled by p42 MAPK and cyclin B/Cdc2 (Fig. 1d). Both of these protein kinases are activated during maturation: p42 MAPK through phosphorylation by MAPK kinase (MEK), and cyclin B/Cdc2 through dephosphorylation by Cdc25. In addition, both kinases seem to be actively maintained in their high-activity, M-phase states. For example, in mature oocytes where p42 MAPK activity is high and unchanging, the two phosphorylation events that activate p42 MAPK still turn over rapidly with half-lives ($t_{1/2}$) of about 5 min (ref. 10). EDTA-quenching experiments suggest that phosphorylation events turn over rapidly ($t_{1/2} < 10$ min) in other members of the MAPK cascade, Mos and MEK, and also in Cdc25 and Wee1 (data not shown). This raises the issue of how these intrinsically reversible signalling proteins can ‘remember’ a transient stimulus and convert it into an irreversible response.

An attractive possibility is that the irreversibility is produced by positive feedback, and indeed numerous positive feedback loops have been identified in the p42 MAPK/Cdc2 network of an oocyte. For example, Mos activates p42 MAPK through the intermediacy of MEK, and active p42 MAPK feeds back to promote the accumulation of Mos (refs 11–13 and Fig. 1d). This protein-synthesis-dependent positive feedback loop is strong enough to allow p42 MAPK to show an all-or-none, bistable response to progesterone or microinjected Mos¹⁴. If protein synthesis is blocked (and thus the positive feedback is blocked), however, the response becomes graded¹⁴.

Other positive feedback loops are also present in the signalling network that culminates in oocyte maturation. Cdc2 promotes activation of the Cdc2 activator Cdc25 (refs 15, 16), and promotes inactivation of the Cdc2 inhibitor Myt1 (ref. 17 and Fig. 1d). The p42 MAPK and Cdc2 systems also interact with each other through positive feedback: p42 MAPK acts positively on Cdc2 by promoting the inactivation of Myt1 (ref. 18), and Cdc2 acts positively on p42 MAPK by promoting the accumulation of Mos (ref. 19 and Fig. 1d). Thus, positive feedback is a recurring theme in oocyte signal transduction, and indeed positive feedback and bistability are important in various biological contexts^{20–23}.

In theory, a system whose positive feedback is strong enough to produce bistability should show either hysteresis (meaning that it is easier to maintain the system in its on state than to toggle the system from off to on) or, if the feedback is particularly strong, irreversibility (Box 1). We considered that the bistable p42 MAPK/Cdc2 system of the oocyte might possess sufficiently strong positive feedback to allow the system to generate this type of self-sustaining, actively maintained irreversibility.

To test this hypothesis, we set out to determine whether hysteresis or irreversibility is apparent in the oocyte’s response to progesterone. This amounts to determining whether the concentration of progesterone needed to induce p42 MAPK phosphorylation and Cdc2 activation differs from that needed to maintain the activities of these kinases. To obtain ‘induction’ stimulus–response curves, we incubated immature oocytes with different concentrations of progesterone, waited until oocyte maturation had reached a plateau, and then measured the percentage of GVBD (%GVBD), p42 MAPK phosphorylation, Cdc2 activation and progesterone binding. Progesterone-treated oocytes showed dose-dependent increases in all of these responses (Fig. 2a–c, ‘induction’), with maximal kinase activation and %GVBD obtained with 600 nM progesterone.

To obtain ‘maintenance’ stimulus–response curves, we incubated oocytes with 600 nM progesterone, waited until GVBD had reached a plateau (5 h in the experiments shown in Fig. 2), washed the oocytes for 10 h to remove progesterone, and then resuspended the washed oocytes in various concentrations of progesterone. After 5 h of further incubation, we monitored %GVBD, p42 MAPK phosphorylation, Cdc2 activation and progesterone binding. Oocytes were found to maintain maximal p42 MAPK and Cdc2 activities after being washed free of progesterone (Fig. 2b, c, ‘maintenance’),

and none of the oocytes lost its white dot (Fig. 2a). The progesterone-binding data argued against the trivial explanation that the irreversibility in the p42 MAPK and Cdc2 responses was simply due to a failure to wash the progesterone away adequately (Fig. 2d). Thus, once oocytes are mature, the continued presence of progesterone seems to be unnecessary. Some 'memory' of the progesterone must be maintained either by the p42 MAPK/Cdc2 system or by signal transducers upstream of this system.

To test whether the p42 MAPK/Cdc2 system itself can generate an irreversible response from a transient stimulus, we used a chimaeric protein of Raf and the oestrogen receptor (Δ Raf:ER), in which an activated Raf1 protein is rendered conditional by fusion to an ER hormone-binding domain²⁴. In oocytes expressing Δ Raf:ER, the steroid hormone oestradiol (which by itself has no effect on maturation, p42 MAPK activation or Cdc2 activation in oocytes) can be used to introduce a stimulus directly into the MAPK cascade, bypassing the progesterone receptor and other upstream signalling proteins.

Oocytes expressing Δ Raf:ER possessed low but detectable Δ Raf:ER

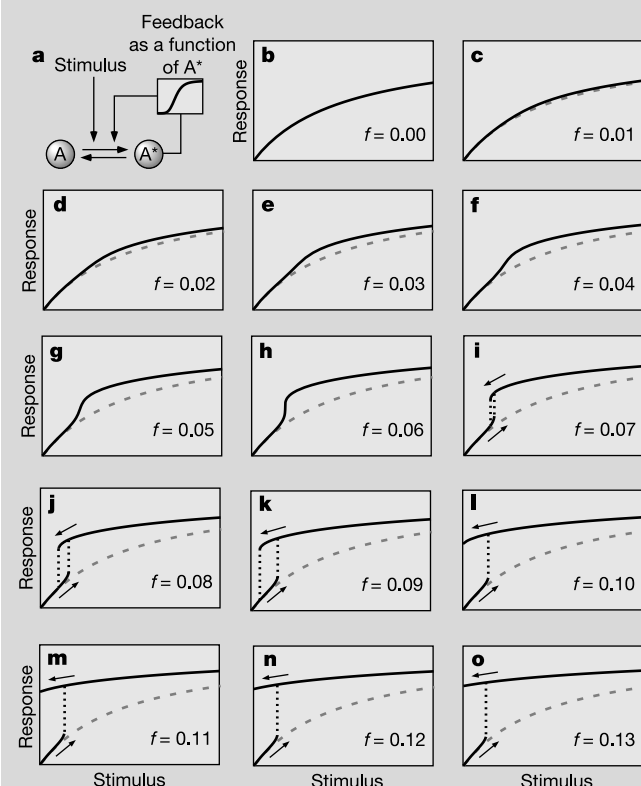
activity in the absence of oestradiol (Fig. 3a, middle, lane 3). In response to oestradiol, there was a prompt 3–5-fold increase in Δ Raf:ER activity (Fig. 3a, middle, lanes 3–5), followed by a slower, additional 3–5-fold increase in Δ Raf:ER protein and activity (Fig. 3a, top and middle, lanes 6–10). The MEK inhibitor PD98059 and the protein synthesis inhibitor cycloheximide blocked the slow increase in Δ Raf:ER protein (data not shown) and activity (Fig. 4a, c), indicating that this component of the Δ Raf:ER activation depends on positive feedback between p42 MAPK and Δ Raf:ER. Within several hours, the oestradiol-induced activation of Δ Raf:ER resulted in complete activation of endogenous p42 MAPK (Fig. 3a, bottom, lanes 9–10).

We used the Δ Raf:ER-expressing oocytes to look for hysteresis or irreversibility in the response to oestradiol, by using the induction versus maintenance strategy described above. When immature oocytes expressing Δ Raf:ER were treated with increasing concentrations of oestradiol, there was a graded, dose-dependent increase in %GVBD and in steady-state Δ Raf:ER activity, p42 MAPK phosphorylation and Cdc2 activity (Fig. 3b–e, 'induction'). The

Box 1

Behaviour of a simple positive feedback loop: sensitivity amplification, bistability, hysteresis and irreversibility

Positive feedback loops have the potential to convert a transient stimulus into a self-sustaining, irreversible response. But irreversibility is not an inevitable consequence of positive feedback, nor is irreversibility the only useful systems-level property that can emerge from systems with positive feedback loops. This is true in the case of complicated positive feedback systems, such as the p42 MAPK/Cdc2 system in oocytes, and is also true (and can be more easily seen and understood) in simpler systems, such as the one shown here (a).



This system consists of a signalling protein that can be reversibly converted between an inactive form (A) and an active one (A*). The

activation reaction is assumed to be regulated in two ways: by an external stimulus (equation (1), first term); and by positive feedback, with a nonlinear Hill equation relationship between the amount of A* produced and the rate of production of more A* (equation (1), second term). The inactivation reaction is assumed to be unregulated; its rate is proportional to the concentration of A* (equation (1), third term). Thus,

$$\frac{d[A^*]}{dt} = \{\text{stimulus} \times ([A_{\text{tot}}] - [A^*]) + f \frac{[A^*]^n}{K^n + [A^*]^n} - k_{\text{inact}}[A^*]\} \quad (1)$$

where n denotes the Hill coefficient, K is the effector concentration for half-maximum response (EC_{50}) for the feedback as a function of $[A^*]$, and f represents the strength of the feedback.

This differential equation was solved numerically (by using Mathematica 2.2.2) to determine the relationship between stimulus and steady-state response ($[A^*]$), assuming that $n = 5$, $K = 1$, $k_{\text{inact}} = 0.01$ and stimulus = 0–1, and assuming a range of values of f . The results are shown in b–o, with the calculated steady-state responses shown as unbroken lines and the discontinuities shown as dotted lines; the no-feedback response (dashed lines) is included for comparison.

When $f = 0$, the response is monostable and the stimulus–response curve is a smooth Michaelian hyperbola (b). As the strength of the feedback increases, the stimulus–response curve acquires a sigmoidal shape (c–h). This occurs because the feedback has been assumed to be cooperative or ultrasensitive. The sigmoidicity makes the response of A* more switch-like (but still monostable).

At $f = 0.07$ (i), the stimulus–response curve splits into two curves: one representing the amount of stimulus needed to induce the system to turn on, the other representing the amount of stimulus needed to maintain the system in its on state. At this point, the system becomes bistable for some values of stimulus (that is, there are two discrete, stable steady states for a single value of stimulus) and the system shows hysteresis, meaning that the dose–response relationship is a loop rather than a curve. The range of stimulus over which the system is bistable and the extent of the hysteresis both increase as f increases (i–k).

Eventually, the feedback becomes strong enough to maintain the system in the on state even when the stimulus is decreased to zero (l–o). At this point, the system may be able to convert a transient stimulus into an essentially irreversible response. But even in a system such as this, stochastic effects still have the potential to make the response reversible^{29,30}.

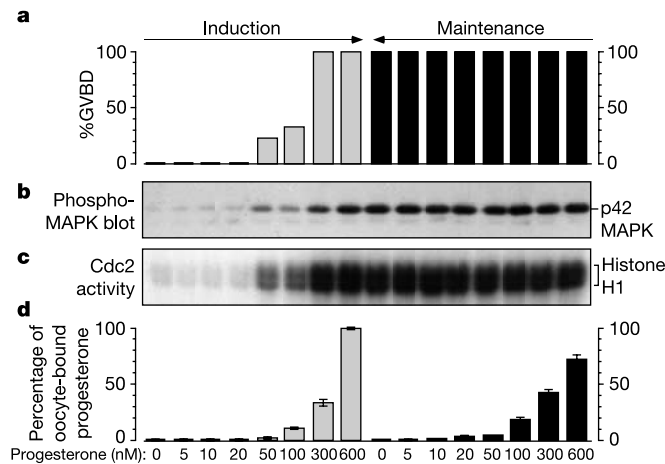


Figure 2 Irreversibility in the biochemical responses of oocytes to progesterone. GVBD (**a**), p42 MAPK phosphorylation (**b**), Cdc2 H1 kinase activity (**c**) and progesterone binding (**d**) were assessed at the end of the induction period and the end of the maintenance period. GVBD, MAPK phosphorylation and Cdc2 activity data are from one of three similar experiments. Progesterone binding data (shown as means $\pm$ s.d.) are from a separate experiment.

kinases remained maximally active after the oestradiol was washed away (Fig. 3c–e, ‘maintenance’), and every oocyte maintained its white spot after washing (Fig. 3b). The trivial explanation of incomplete removal of oestradiol did not seem to account for the observed irreversibility (Fig. 3f). Thus, the p42 MAPK/Cdc2 system itself can convert a transient stimulus into an irreversible activation response; that is, the system can generate a long-term ‘memory’ of a transient differentiation stimulus.

To determine whether positive feedback is essential for maintaining this memory, we made use of three ways of blocking feedback from p42 MAPK to Mos. The first feedback inhibitor was cycloheximide. If oocytes are provided with a sufficiently strong maturation stimulus, such as microinjected Mos or cyclin A, protein synthesis is not essential for maturation^{25,26}; however, protein synthesis is essential for the feedback from p42 MAPK to Mos^{11–13}, and possibly for other interconnected positive feedback loops²⁷. Thus, we examined whether cycloheximide would convert the observed irreversible oestradiol response into a reversible one. In the absence of cycloheximide, oestradiol again caused marked increases in the steady-state activities of Δ Raf:ER, p42 MAPK and Cdc2, and these responses were undiminished after washing (Fig. 4a). In the presence of cycloheximide, the responses of Δ Raf:ER and p42 MAPK to oestradiol were blunted, and the response of Cdc2 was markedly diminished (Fig. 4a). In addition, the responses were no longer irreversible: after the oestradiol was washed away, the activity of Δ Raf:ER, the phosphorylation of p42 MAPK, and the low activation of Cdc2 all decreased to near-basal levels (Fig. 4a). Thus, protein synthesis, which is essential for feedback from p42 MAPK to Mos, is also required for the irreversibility of the oestradiol responses.

The second feedback inhibitor was a morpholino Mos antisense oligonucleotide (Mos-AS), which blocks progesterone-induced Mos synthesis without globally abolishing protein synthesis²⁸. Mos-AS blocked oestradiol-induced Mos synthesis and also abolished the irreversibility of Δ Raf:ER activation and p42 MAPK phosphorylation (Fig. 4b). The low Cdc2 activation induced by oestradiol also became reversible (Fig. 4b). Thus, Mos synthesis is required for the observed irreversibility. Finally, we examined the effects of the MEK inhibitor PD98059, which blocks both downstream effects and feedback effects that depend on MEK activation. PD98059 rendered oestradiol-induced Δ Raf:ER activation revers-

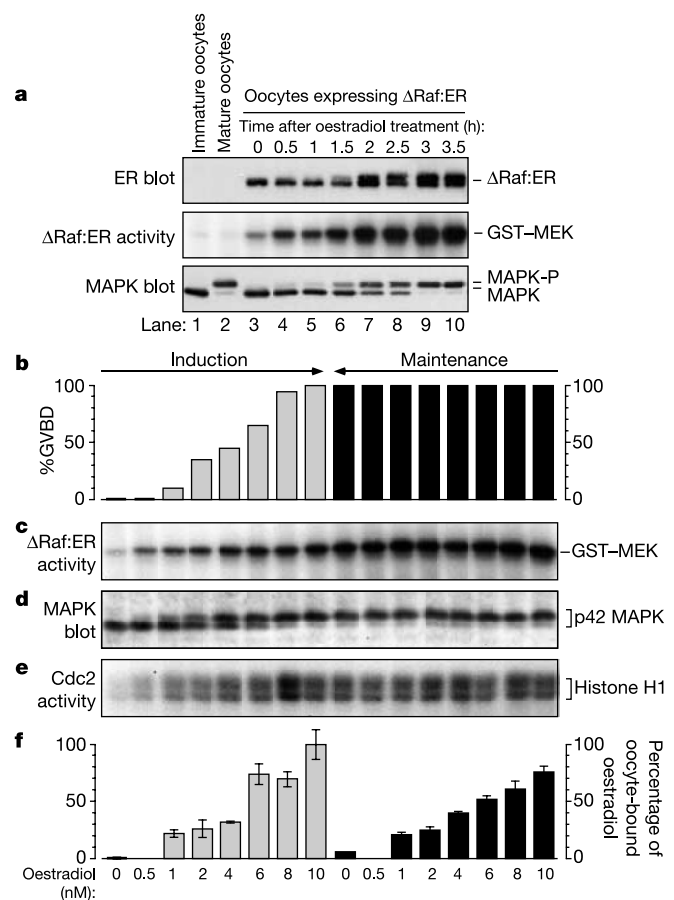


Figure 3 Irreversibility in the biochemical responses of oocytes expressing Δ Raf:ER to oestradiol. **a**, Time course of oestradiol-induced Δ Raf:ER and p42 MAPK activation, assessed by ER immunoblot analysis, Δ Raf:ER immune complex kinase assay, and p42 MAPK immunoblot analysis. **b–f**, Steady-state responses. GVBD (**b**), Δ Raf:ER activity (**c**), p42 MAPK phosphorylation (**d**), Cdc2 H1 kinase activity (**e**) and oestradiol binding (**f**) were assessed at the end of the induction period and the end of the maintenance period. GVBD, Δ Raf:ER activity, p42 MAPK phosphorylation and Cdc2 activity data are from one of three similar experiments. Oestradiol binding data (shown as means $\pm$ s.d.) are from a separate experiment.

ible and also blocked oestradiol-induced p42 MAPK and Cdc2 activation (Fig. 4c).

Taken together, these findings rule out the possibility that an inability to wash away Δ Raf:ER-bound oestradiol, or an intrinsic defect in the capacity of Δ Raf:ER to be inactivated, can account for the irreversible p42 MAPK and Cdc2 responses seen in Figs 3 and 4. Instead, the irreversibility seems to be actively generated by a bistable, positive feedback system, and compromising the feedback abolishes the irreversibility.

The idea that the irreversibility of the differentiated state might be maintained by feedback loops that generate self-sustaining patterns of gene expression dates back more than 40 years (ref. 4), and the recognition that feedback-enforced patterns of protein activity might also produce a systems-level ‘memory’ of transient stimuli dates back almost 20 years (ref. 5). Studies of artificial bistable gene expression systems in *Escherichia coli* and *Saccharomyces cerevisiae* have shown that bistable systems can in fact function as memory modules, but have also shown that stochastic effects sometimes cause two bistable steady states to equilibrate with each other, undermining any memory^{29,30}. Our work provides experimental evidence that in a physiological process of cell fate induction, *Xenopus* oocyte maturation, a bistable signalling system

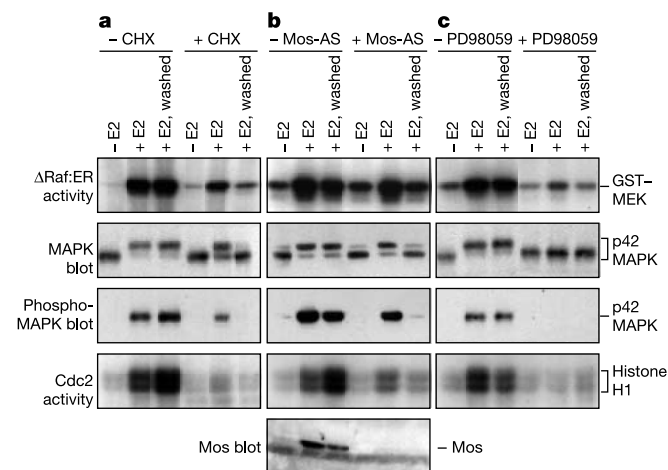


Figure 4 Positive feedback is required for irreversible biochemical responses. Oocytes expressing Δ Raf:ER were treated by an 18-h incubation with no added oestradiol (– E2); an 18-h incubation with 10 nM oestradiol (+ E2); or a 2-h incubation with 10 nM oestradiol, followed by a 16-h wash (+ E2, washed). **a**, A sample of oocytes was treated with cycloheximide (CHX, 100 μ g ml^{–1}) to block protein synthesis and positive feedback. **b**, A sample of oocytes was injected with a morpholino Mos antisense oligonucleotide (100 ng per oocyte) to block Mos synthesis. **c**, A sample of oocytes was treated with the MEK inhibitor PD98059 (100 μ M) to block MEK-mediated positive feedback and downstream events.

converts a transient stimulus into a reliable, self-sustaining, effectively irreversible pattern of protein activities. It will be interesting to see how common these bistable memory modules are in cell signalling. □

Methods

Oocytes, expression constructs and reagents

Xenopus laevis oocytes were defolliculated by collagenase treatment and stage VI oocytes were manually sorted. We prepared oocyte lysates as described¹⁰.

The high-activity DD form of Δ Raf:ER (in pBabepuro3)²⁴ was a gift from M. McMahon and was subcloned into pSP64(polyA). Δ Raf:ER mRNA was transcribed *in vitro* using an Sp6 transcription kit (Ambion), and the RNA concentration was determined by gel electrophoresis and staining. Δ Raf:ER mRNA (5 ng) was microinjected with a Nanoinjector (Drummond Scientific), and oocytes were left for protein expression for 2 h before oestradiol was added. We purchased the Mos antisense morpholino oligonucleotide 5′-AAGGCATTGCTGTGTGACTCGTGA-3′ (ref. 28) from Gene Tools LLC.

Oestradiol, progesterone and cycloheximide were from Sigma, and [³H]oestradiol and [¹⁴C]progesterone were from New England Nuclear. The MEK inhibitor PD98059 was from Calbiochem.

SDS-PAGE and immunoblotting

For most assays, samples were resolved by electrophoresis on 10% low-bis polyacrylamide gels (acrylamide:bisacrylamide 100:1). For the Cdc2/cyclin B histone H1 kinase assays, we used 12.5% polyacrylamide gels (acrylamide:bisacrylamide 29:1). Gels were transferred to poly(vinylidene difluoride) blotting membranes, which were then blocked with 3% nonfat milk in Tris-buffered saline (150 mM NaCl and 20 mM Tris; pH 7.6) and incubated for 2 h with a 1:1,000 dilution of one of the following primary antibodies: anti-MAPK (DC3, raised in our own laboratory), anti-phospho-MAPK (9106, New England Biolabs), anti-ER (SC-543, Santa Cruz Biotechnology) or a 1:500 dilution of anti-Mos (SC-86, Santa Cruz Biotechnology). Blots were washed three times with Tris-buffered saline plus 0.1% Tween 20 and probed with alkaline-phosphatase-conjugated secondary antibody for detection by CDP-Star (Tropix-Perkin Elmer). For reprobing, blots were stripped in 100 ml of stripping buffer (100 mM Tris-HCl, pH 7.4, 100 mM β -mercaptoethanol and 2% SDS) at 70 °C for 40 min.

Kinase assays

We prepared immune complexes of Δ Raf:ER by incubating cell lysate with 30 μ l of protein-A-Sepharose (Sigma) pre-coated with 1 μ l of anti-ER antibody (SC-543, Santa Cruz Biotechnology). The immunoprecipitates were washed once with lysis buffer and once with kinase buffer (20 mM HEPES, 10 mM MgCl₂ and 1 mM MnCl₂; pH 7.4). Raf kinase reactions were done at 30 °C for 30 min in 40 μ l of kinase buffer with 1 mM dithiothreitol, 1 μ M ATP and 10 μ Ci [³²P]ATP (Amersham) with 50 ng of purified recombinant GST-MEK (UBI) as a substrate. The reaction was stopped by boiling in sample buffer and resolved by electrophoresis on a 10% 100:1 polyacrylamide gel.

For the histone H1 kinase assay, oocytes were lysed in a volume of 10 μ l per oocyte, and 2 μ l of cleared lysate was added to 8 μ l of EB buffer (80 mM β -glycerophosphate, 20 mM EGTA and 15 mM MgCl₂; pH 7.3) and mixed with an equal volume (10 μ l) of reaction mixture (6.73 μ l of 2 $\times$ H1 kinase buffer, 0.02 μ l of 0.2 mM ATP, 1.0 μ l of 10 mg ml^{–1} histone H1, 2.0 μ l of 100 μ M protein kinase inhibitor (Santa Cruz Biotechnology) and 0.25 μ l of 10 mCi ml^{–1} [³²P]ATP). Reactions were carried out at 30 °C for 15 min and terminated by the addition of 6 $\times$ sample buffer (4 μ l) followed by boiling.

Washing procedure and scintillation counting

Oocytes incubated with progesterone or oestradiol were washed free of the steroids by a series of ten 1-h washes in 50 ml of OR2 solution. For some experiments, cycloheximide or MEK inhibitor (PD98059) was included in both the incubation and the washing buffer. Pre- and post-wash binding of progesterone and oestradiol was determined by the inclusion of radiolabelled tracers ([¹⁴C]progesterone, final specific activity 83 μ Ci μ mol^{–1}, and [³H]oestradiol, final specific activity 25,000 μ Ci μ mol^{–1}) and quantified by scintillation counting.

Received 9 April; accepted 18 September 2003; doi:10.1038/nature02089.

- Abrieu, A., Doree, M. & Fisher, D. The interplay between cyclin-B-Cdc2 kinase (MPF) and MAP kinase during maturation of oocytes. *J. Cell Sci.* **114**, 257–267 (2001).
- Nebreda, A. R. & Ferby, I. Regulation of the meiotic cell cycle in oocytes. *Curr. Opin. Cell Biol.* **12**, 666–675 (2000).
- Ferrell, J. E. Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Curr. Opin. Cell Biol.* **14**, 140–148 (2002).
- Monod, J. & Jacob, F. General conclusions: teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harb. Symp. Quant. Biol.* **26**, 389–401 (1961).
- Lisman, J. E. A mechanism for memory storage insensitive to molecular turnover: a bistable autophosphorylating kinase. *Proc. Natl Acad. Sci. USA* **82**, 3055–3057 (1985).
- Thomas, R. & Kaufman, M. Multistationarity, the basis of cell differentiation and memory. I. Structural conditions of multistationarity and other nontrivial behavior. *Chaos* **11**, 170–179 (2001).
- Dettlaff, T. A. Action of actinomycin and puromycin upon frog oocyte maturation. *J. Embryol. Exp. Morphol.* **16**, 183–195 (1966).
- Schuetz, A. W. Action of hormones on germinal vesicle breakdown in frog (*Rana pipiens*) oocytes. *J. Exp. Zool.* **166**, 347–354 (1967).
- Smith, L. D., Ecker, R. E. & Subtelny, S. *In vitro* induction of physiological maturation in *Rana pipiens* oocytes removed from their ovarian follicles. *Dev. Biol.* **17**, 627–643 (1968).
- Sahaskey, M. L. & Ferrell, J. E. Jr Distinct, constitutively active MAPK phosphatases function in *Xenopus* oocytes: implications for p42 MAPK regulation *in vivo*. *Mol. Biol. Cell* **10**, 3729–3743 (1999).
- Matten, W. T., Copeland, T. D., Ahn, N. G. & Vande Woude, G. F. Positive feedback between MAP kinase and Mos during *Xenopus* oocyte maturation. *Dev. Biol.* **179**, 485–492 (1996).
- Roy, L. M. *et al.* Mos proto-oncogene function during oocyte maturation in *Xenopus*. *Oncogene* **12**, 2203–2211 (1996).
- Gotoh, Y., Masuyama, N., Dell, K., Shirakabe, K. & Nishida, E. Initiation of *Xenopus* oocyte maturation by activation of the mitogen-activated protein kinase cascade. *J. Biol. Chem.* **270**, 25898–25904 (1995).
- Ferrell, J. E. Jr & Machleder, E. M. The biochemical basis of an all-or-none cell fate switch in *Xenopus* oocytes. *Science* **280**, 895–898 (1998).
- Kumagai, A. & Dunphy, W. G. Regulation of the cdc25 protein during the cell cycle in *Xenopus* extracts. *Cell* **70**, 139–151 (1992).
- Hoffmann, L., Clarke, P. R., Marcote, M. J., Karsenti, E. & Draetta, G. Phosphorylation and activation of human cdc25-C by cdc2–cyclin B and its involvement in the self-amplification of MPF at mitosis. *EMBO J.* **12**, 53–63 (1993).
- Mueller, P. R., Coleman, T. R., Kumagai, A. & Dunphy, W. G. Myt1: a membrane-associated inhibitory kinase that phosphorylates Cdc2 on both threonine-14 and tyrosine-15. *Science* **270**, 86–90 (1995).
- Palmer, A., Gavin, A. C. & Nebreda, A. R. A link between MAP kinase and p34^{cdc2}/cyclin B during oocyte maturation: p90^{rsk} phosphorylates and inactivates the p34^{cdc2} inhibitory kinase Myt1. *EMBO J.* **17**, 5037–5047 (1998).
- Nebreda, A. R., Gannon, J. V. & Hunt, T. Newly synthesized protein(s) must associate with p34^{cdc2} to activate MAP kinase and MPF during progesterone-induced maturation of *Xenopus* oocytes. *EMBO J.* **14**, 5597–5607 (1995).
- Bhalla, U. S., Ram, P. T. & Iyengar, R. MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signalling network. *Science* **297**, 1018–1023 (2002).
- Sha, W. *et al.* Hysteresis drives cell-cycle transitions in *Xenopus laevis* egg extracts. *Proc. Natl Acad. Sci. USA* **100**, 975–980 (2003).
- Pomeroy, J. R., Sontag, E. D. & Ferrell, J. E. Jr Building a cell cycle oscillator: hysteresis and bistability in the activation of Cdc2. *Nature Cell Biol.* **5**, 346–351 (2003).
- Bagowski, C. P. & Ferrell, J. E. Bistability in the JNK cascade. *Curr. Biol.* **11**, 1176–1182 (2001).
- Bosch, E., Cherwinski, H., Peterson, D. & McMahon, M. Mutations of critical amino acids affect the biological and biochemical properties of oncogenic A-Raf and Raf-1. *Oncogene* **15**, 1021–1033 (1997).
- Yew, N., Mellini, M. L. & Vande Woude, G. F. Meiotic initiation by the *mos* protein in *Xenopus*. *Nature* **355**, 649–652 (1992).
- Roy, L. M. *et al.* Activation of p34^{cdc2} kinase by cyclin A. *J. Cell Biol.* **113**, 507–514 (1991).
- Groisman, I., Jung, M.-Y., Sarkissian, M., Cao, Q. & Richter, J. D. Translational control of the embryonic cell cycle. *Cell* **109**, 473–483 (2002).
- Dupre, A., Jessus, C., Ozon, R. & Haccard, O. Mos is not required for the initiation of meiotic maturation in *Xenopus* oocytes. *EMBO J.* **21**, 4026–4036 (2002).
- Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403**, 339–342 (2000).
- Becskei, A., Seraphin, B. & Serrano, L. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *EMBO J.* **20**, 2528–2535 (2001).

Acknowledgements We thank M. McMahon for the Δ Raf:ER constructs, and K. Cimprich and members of the Ferrell laboratory for discussions and comments on the manuscript. This work was supported by a grant from the NIH.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.E.F. (james.ferrell@stanford.edu).

Structure and nucleic-acid binding of the *Drosophila* Argonaute 2 PAZ domain

Andreas Lingel*, Bernd Simon*, Elisa Izaurralde & Michael Sattler

European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

* These authors contributed equally to this work

RNA interference is a conserved mechanism that regulates gene expression in response to the presence of double-stranded (ds)RNAs^{1,2}. The RNase III-like enzyme Dicer first cleaves dsRNA into 21–23-nucleotide small interfering RNAs (siRNAs)^{3–6}. In the effector step, the multimeric RNA-induced silencing complex (RISC) identifies messenger RNAs homologous to the siRNAs and promotes their degradation^{3,7}. The Argonaute 2 protein (Ago2) is a critical component of RISC^{8,9}. Both Argonaute and Dicer family proteins contain a common PAZ domain whose function is unknown¹⁰. Here we present the three-dimensional nuclear magnetic resonance structure of the *Drosophila melanogaster* Ago2 PAZ domain. This domain adopts a nucleic-acid-binding fold that is stabilized by conserved hydrophobic residues. The nucleic-acid-binding patch is located in a cleft between the surface of a central β -barrel and a conserved module comprising strands β 3, β 4 and helix α 3. Because critical structural residues and the binding surface are conserved, we suggest that PAZ domains in all members of the Argonaute and Dicer families adopt a similar fold with nucleic-acid binding function, and that this plays an important part in gene silencing.

Ago2 is a member of the PPD family of proteins (PAZ and Piwi domains), which are characterized by an amino-terminal PAZ domain (named after the Piwi, Argonaute and Zwiller/Pinhead proteins) and a carboxy-terminal Piwi domain^{10,11}. The PPD family comprises highly basic proteins with relative molecular masses of approximately 100,000 (M_r 100K) that are present in metazoa and fungi but not in budding yeast. Members of this protein family have essential roles in RNA interference (RNAi) and related gene-silencing processes in several organisms, and have also been implicated in cell-fate determination¹². The molecular function of these proteins is unknown. The Piwi domain is highly conserved and exists also in prokaryotes¹⁰. The less conserved 110-residue PAZ domain has only been identified in eukaryotes¹⁰. It is found exclusively in the Argonaute and Dicer protein families.

The solution structure of the Ago2 PAZ domain (Fig. 1 and Table 1) comprises a central five-stranded open β -barrel (strands β 1, β 2, β 5– β 7), an N-terminal helical region (helices α 1, α 2) and a conserved approximately 35-residue module (comprising β 3, β 4 and α 3) between strands β 2 and β 5. The N and C termini of the PAZ fold (residues 1–120) form a small antiparallel β -sheet (strands β 0, β 7) and thus are in close spatial proximity, as expected for an independent structural domain. Notably, the β 5– β 6 loop has an

approximately 30-residue insertion in PAZ domains of the Dicer proteins (Supplementary Fig. S1), which could modulate their function. On the basis of nuclear magnetic resonance (NMR) relaxation measurements the PAZ domain is monomeric in solution (Supplementary Fig. S2).

The structure of the PAZ domain is stabilized by conserved hydrophobic residues, suggesting a similar fold for all members of the family. The hydrophobic core of the central β -barrel is formed by Ile 40, Val 42, Tyr 44, Tyr 57, Val 59, Leu 62, Leu 100, Val 102, Ile 109, Leu 111 and Ile 113. Hydrophobic side chains are conserved in the interface between the β 3, β 4, α 3 module (Ile 81, Tyr 84, Phe 85, Tyr 90) and the central β -barrel (Tyr 44, Cys 99, Pro 112, Leu 115). The strands β 6 and β 7 and the connecting 3_{10} helix are highly conserved among PAZ domains (Supplementary Figs S1 and S3a). The 3_{10} helix contains the invariable Glu 114 residue, which contacts Tyr 90 and Phe 94 and is positioned to form a hydrogen bond with the Lys 93 backbone amide. This leads to tight packing between the α 3– β 5 loop and β 6/ β 7, additionally stabilizing the PAZ domain fold. Consistently, a Glu 114 to Lys mutation yields insoluble protein upon expression in *Escherichia coli* (data not shown). In a genetic screen in *Caenorhabditis elegans*, worms deficient in RNAi were isolated in which the corresponding Glu was mutated to Lys in the Ago2 homologue RDE-1 (ref. 13). Our structural analysis suggests that this mutation probably destabilizes the fold of the PAZ domain, resulting in a non-functional RDE-1 protein.

A surface plot coloured by the degree of sequence conservation among PAZ domains (Fig. 2a) reveals a region comprising the β 3, β 4, α 3 module and the central β -barrel, which together form a clamp-like structure (Figs 1b and 2a). The presence of aromatic and positively charged residues at this surface (Fig. 2b) may indicate a conserved function in the binding of negatively charged ligands, such as nucleic acids. A further hint at a potential role of PAZ domains in nucleic-acid binding comes from the five-stranded

Table 1 Structural statistics for the Ago2 PAZ domain

Statistics	(SA)*	(SA ^{watref})*
r.m.s. deviation (Å) from experimental distance restraints†		
Unambiguous/ambiguous (3,370/46)	0.0078 ± 0.0009	0.014 ± 0.001
Hydrogen bonds (18 × 2)	0.021 ± 0.002	0.042 ± 0.003
R.m.s. deviation (°) from experimental torsion restraints‡		
Dihedral angles (139 ϕ , 105 ψ)	0.44 ± 0.06	0.68 ± 0.1
R.m.s. deviation (Hz) from residual dipolar coupling restraints§		
H ^N –N (32)	0.62 ± 0.09	1.00 ± 0.11
Coordinate precision (Å) residues 4–120		
N, C α , C' (secondary structure elements)	0.40 ± 0.04	0.51 ± 0.09
N, C α , C'	0.49 ± 0.08	0.59 ± 0.08
All heavy atoms	0.94 ± 0.07	1.04 ± 0.06
Structural quality¶		
Bad contacts	2.3 ± 0.8	0.3 ± 0.5
Ramachandran plot		
Per cent in most favoured region	83.7 ± 1.9	89.9 ± 2.5
Per cent in additionally allowed region	14.8 ± 2.0	8.9 ± 2.5

* (SA) is an ensemble of ten lowest-energy solution structures of the Ago2 PAZ domain (out of 100 calculated). The CNS E_{repl} function was used to simulate van der Waals interactions with an energy constant of 25.0 kcal mol⁻¹ Å⁻⁴ using 'PROLSQ' van der Waals radii. Root mean square (r.m.s.) deviations for bond lengths, bond angles and improper dihedral angles are 0.00156 ± 0.00004 Å, 0.327 ± 0.005° and 0.256 ± 0.014°, respectively. 1 kcal = 4.18 kJ. For (SA^{watref}), the ensemble of (SA) structures was refined in a shell of water as described²⁴.

†Distance restraints were used with a soft square-well potential using an energy constant of 50 kcal mol⁻¹ Å⁻². For hydrogen bonds, distance restraints with bounds of 1.8–2.3 Å (H–O) and 2.8–3.3 Å (N–O) were derived for slow-exchanging amide protons. No distance restraint was violated by more than 0.3 Å in any of the final (SA) structures.

‡Dihedral angle restraints, derived from ³J(H^N,H α) coupling constants and TALOS²¹, were applied to ϕ and ψ using energy constants of 200 kcal mol⁻¹ rad⁻². No dihedral angle restraint was violated by more than 5°.

§Residual dipolar couplings were applied with a final energy constant of 0.3 kcal mol⁻¹ Hz⁻² for an alignment tensor with an axial component of 16.2 Hz and a rhombicity of 0.43.

||Coordinate precision is given as the cartesian coordinate r.m.s. deviation of the ten lowest-energy structures with respect to their mean structure.

¶Structural quality was analysed using PROCHECK²⁵.

β -barrel comprising β 1, β 2 and β 5– β 7. This region has some resemblance with the OB (oligonucleotide and oligosaccharide-binding)-fold, a structure that is implicated most frequently in single-stranded nucleic-acid binding¹⁴. The central β -barrel of the PAZ domain is a topological variation of this fold, where the β 4 strand in the OB-fold is replaced by an N-terminal strand (β 1) in the PAZ domain (Fig. 1c). In contrast to OB-fold proteins, where the β -barrel is solvent exposed, the central β -barrel in the PAZ domain faces the conserved β 3, β 4, α 3 module, together forming the clamp-like structure.

On the basis of the role of Ago2 in RNAi and the existence of an OB-fold-related structural element, we tested whether the PAZ domain may indeed bind nucleic acids. Using NMR titration experiments we found that the Ago2 PAZ domain binds to single-stranded 21-nucleotide sense or antisense siRNAs (siRNA-1s and siRNA-1as, Supplementary Table S1). Chemical shift perturbations on addition of the siRNA (Fig. 2c) show that the RNA is bound in the cleft between the β -sheet surface (β 2, β 5) and the β 3, β 4, α 3 module. The RNA-binding region involves highly conserved residues, specifically Arg 55 (β 2), Phe 72 (β 3), Thr 80 (β 4), Val 102 (β 5) and Ile 81 (α 3), but also the less conserved Phe 50 in the β 1– β 2 loop

(Figs 1b and 2d).

To identify key residues involved in RNA binding, we performed a mutational analysis of solvent-exposed residues (Phe 50, Arg 55, Arg 58, Arg 64 and Phe 72). The nucleic-acid-binding activity of the PAZ mutants was tested in ultraviolet cross-linking experiments using recombinant proteins and a [³²P]-end-labelled siRNA (siRNA-2s). The wild-type PAZ domain, the Arg64Ala mutant and the Arg55Glu/Arg58Glu double mutant were able to form a photoinduced product (Fig. 3a, asterisk). The proteins cross-linked to the siRNA migrate with a lower mobility in SDS–polyacrylamide gel electrophoresis (PAGE) owing to their increased size (the protein–RNA complexes were not digested with RNase), and were rendered radioactive by the cross-linked RNA (Fig. 3a). Mutation of Phe 72 or Phe 50 to Ala strongly reduced siRNA-2s binding without disrupting the fold of the PAZ domain (based on NMR spectra; data not shown). Consistently, no chemical shift changes are observed in the NMR spectra of the Phe72Ala mutant PAZ domain at equimolar RNA concentration (data not shown). Phe 50 and Phe 72 are located on opposite faces of the clamp in the PAZ domain structure (Figs 1b and 2d). Our results identify Phe 50 and Phe 72 as critical residues for RNA binding.

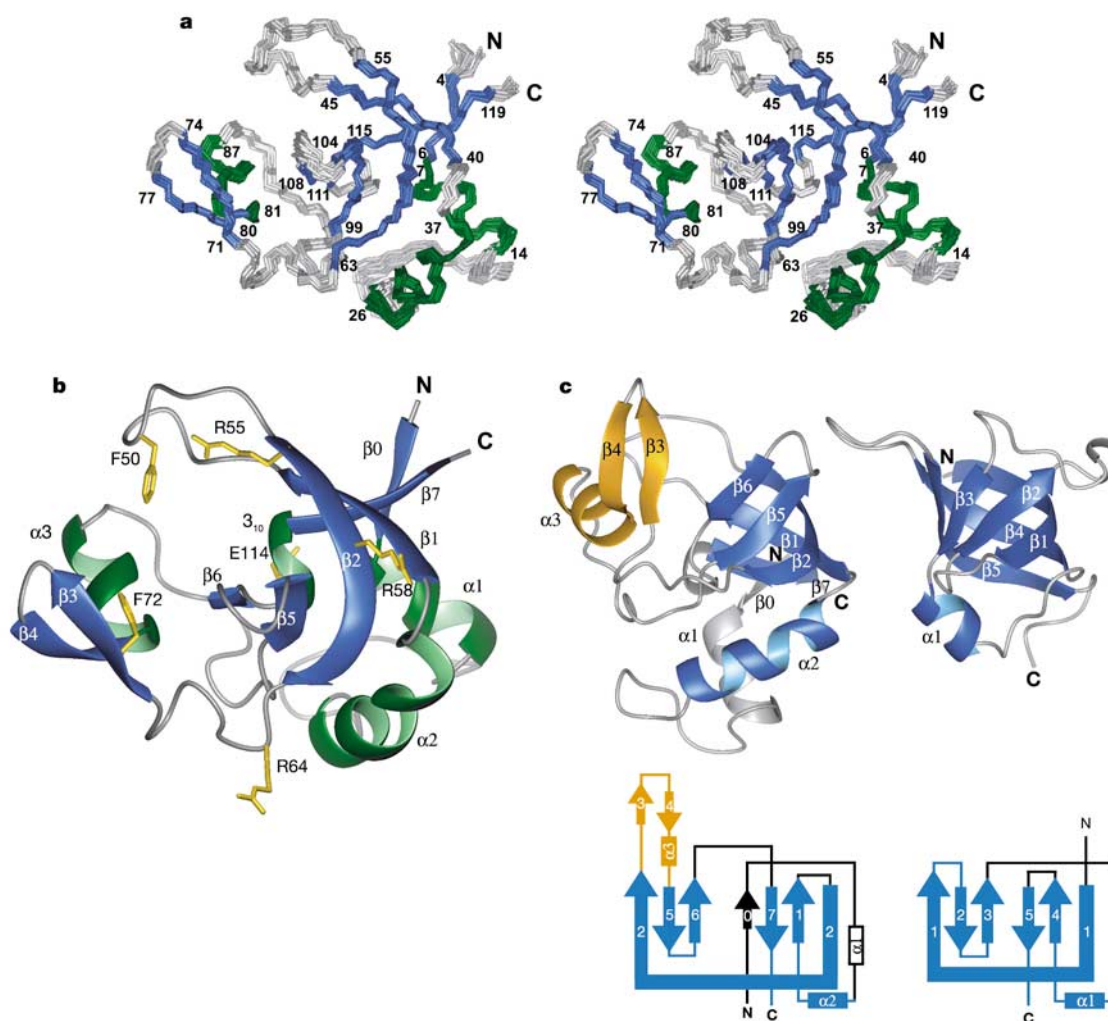


Figure 1 Structure of the Ago2 PAZ domain from *D. melanogaster*. **a**, Stereoview of the ensemble of 15 lowest-energy NMR structures. Helices are shown in green, β -strands in blue. Secondary structure elements are labelled by residue numbers (compare with Supplementary Fig. S1). **b**, Ribbon representation of the Ago2 PAZ domain. Secondary structure elements are labelled. The side chains of residues that were mutated are

annotated and coloured in gold. **c**, Comparison of the central five-stranded β -barrel of the PAZ domain to a canonical OB-fold in the *E. coli* Rho protein³⁰. Related secondary structure elements are coloured in blue. The inserted β 3, β 4, α 3 module of the PAZ domain is shown in orange. The respective topology of the two β -barrels is indicated schematically at the bottom.

The dissociation constant for the interaction of the PAZ domain with single-stranded siRNA derived from the NMR titrations is about 20 μ M. We next tested whether this domain also binds other nucleic acids using the ultraviolet cross-linking assay in the presence of increasing amounts of unlabelled nucleic acid competitors (Fig. 3b–d; see also Supplementary Table S1). The sense or antisense strand of siRNA-1 (siRNA-1s and siRNA-1as) as well as the siRNA-1 duplex (siRNA-1ds) containing two-nucleotide 3' overhangs compete with the formation of [32 P]-labelled photoinduced product as efficiently as unlabelled siRNA-2s (Fig. 3b, c). Similar results were obtained with other siRNAs (not shown), indicating that binding is independent of the sequence. Consistently, in NMR-binding experiments chemical shift perturbations observed on addition of siRNA-1ds to the PAZ domain indicate that it binds with an affinity comparable to siRNA-1s (Supplementary Fig. S4). The 5' phosphate that is present in siRNAs is not specifically recognized because siRNA-1s lacking a 5' monophosphate introduces virtually identical chemical shift changes in the NMR spectra, indicating similar affinity (Supplementary Fig. S4). In contrast, the PAZ domain binds single nucleotides with an approximately 30-fold reduced affinity.

To determine whether the PAZ domain recognizes dsRNA or

mainly the two-nucleotide 3' overhang in siRNA duplexes, we tested binding to two small RNA hairpins with blunt ends (H1-RNA and T1-RNA, Supplementary Table S1). Both RNAs compete with binding of [32 P]-labelled siRNA-2s in the ultraviolet cross-linking experiments, but only when added at twofold higher concentration relative to siRNA-1s (Fig. 3e). These results are confirmed by NMR, where no spectral changes are observed on equimolar addition of T1-RNA or a blunt-ended siRNA-1ds to the PAZ domain (Supplementary Fig. S4), and very few chemical shift changes are observed only at threefold molar excess. We conclude that the PAZ domain has reduced affinity for blunt-ended dsRNA molecules.

We also tested the binding of DNA to the PAZ domain (Fig. 3a, d). Single-stranded 21-nucleotide DNA oligonucleotides corresponding to siRNA-1 (DNA-1, Supplementary Table S1) exhibit a similar relative affinity compared to the single-stranded siRNA-1 (Fig. 3d). Interestingly, the DNA-1ds (Fig. 3d) and other double-stranded DNA molecules are also recognized independently of the presence of a 3' single-stranded overhang (Supplementary Fig. S4 and data not shown). The finding that the PAZ domain also binds single-stranded or double-stranded DNA suggests that it could have a dual function in regulating gene expression by binding RNA or

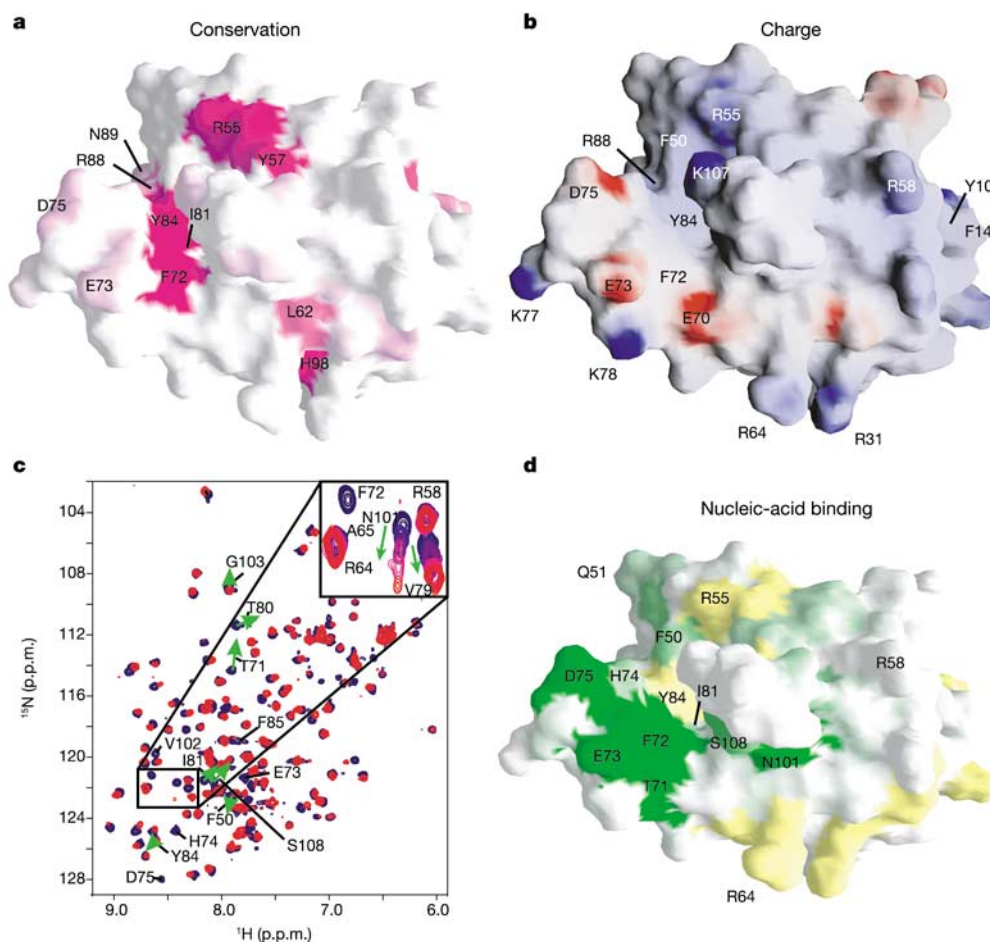


Figure 2 Nucleic-acid-binding surface of the Ago2 PAZ domain. **a**, Surface representation of the PAZ domain coloured by sequence conservation (compare with Supplementary Fig. S1). High to low sequence conservation is coloured from magenta to white, respectively. **b**, Surface representation of the PAZ domain coloured by electrostatic charge. White, blue and red corresponds to neutral, positive and negative electrostatic potential, respectively. **c**, NMR titration of the PAZ domain with a single-stranded siRNA (siRNA-1as). The ^1H - ^{15}N correlation spectrum of the free protein and at sixfold molar excess of siRNA is shown in blue and red, respectively. Green arrows indicate the direction

of peak movements. The insert shows additional spectra at 0.5-, 1-, 1.5- and 3-fold molar excess of RNA. p.p.m., parts per million. **d**, RNA-binding site on the PAZ domain based on the NMR titration in **c**. Prolines or residues that could not be analysed owing to signal overlap are shown in yellow; residues where signals disappear at 1:0.5 molar protein:RNA ratio are coloured green. Other residues are coloured with a gradient from green to white according to the degree of chemical shift changes. In all panels the same orientation as in Fig. 1b is shown.

DNA. This possibility is consistent with the recent observation of a link between the RNAi machinery and regulation of gene expression at the level of heterochromatin¹⁵.

Our data demonstrate that the Ago2 PAZ domain binds nucleic acids in a cleft between a central β -barrel and a highly conserved module comprising β 3, β 4 and α 3. Full-length Ago2 (lacking the N-terminal poly-glutamine region) exhibits a specificity for nucleic acids similar to that of the Ago2 PAZ domain alone (Supplementary Fig. S5), suggesting that the PAZ domain provides a major contribution for nucleic acid recognition. Amino acids important for the

fold and most of the residues in the nucleic-acid-binding site are conserved among PAZ domains, suggesting a related function. In fact, we found that the PAZ domains of *D. melanogaster* Ago1 and Piwi, and human Piwi and eIF2C1 bind nucleic acids (data not shown; see also the accompanying paper for the structure of PAZ from *D. melanogaster* Ago1, ref. 16). Furthermore, RNAi deficiency in nematode worms correlates with a mutation in the PAZ domain of RDE-1, which probably disrupts the PAZ fold *in vivo*. Taken together, these data indicate that the PAZ domain has an essential role in RNAi.

Even though the binding affinity is relatively weak, within the RISC complex the PAZ domain could facilitate recognition of the siRNAs that have been processed from larger dsRNA precursors by the Dicer enzyme. Importantly, recognition of specific RNA topology, namely dsRNA with 3' single-stranded overhangs, rather than RNA sequence *per se* might improve further the efficiency of RNAi. This is fully consistent with the observation that siRNAs processed by Dicer contain 3' single-stranded overhangs^{3,5,17}, and that such siRNAs are more efficient at gene silencing than purely double-stranded siRNAs^{9,18}. Alternatively, the PAZ domain might be involved in binding RNA or DNA sequences that are targeted by RISC for gene silencing. The existence of multiple Argonaute proteins in many organisms suggests that these proteins may assemble in distinct RISC complexes having specific functions. The PAZ domain might contribute to the specificity of these complexes by discriminating between nucleic acid effectors that mediate post-transcriptional gene silencing. □

Methods

Sample preparation

The PAZ domain of the *D. melanogaster* Ago2 protein (TrEMBL entry Q9VUQ5; Met 605–Ser 743, renumbered to residues 5–143 in this manuscript) was amplified from a S2 cell complementary DNA library and cloned into pETM60 vector, derived from pET24-d (Novagen). The protein was expressed in the *E. coli* strain BL21(DE3) at room temperature. For labelling of the protein uniformly with ¹⁵N/¹³C or ¹⁵N, cells were grown in M9 minimal medium supplemented with ¹⁵NH₄Cl with or without ¹³C₆-glucose. Cell lysates were subjected to affinity chromatography purification using Ni-NTA Superflow material (Qiagen), followed by cleavage of the tag with TEV-protease overnight. After a second affinity purification step, the protein was purified to homogeneity by gel filtration. The purity and the molecular mass of the resulting protein, consisting of the cloned sequence plus additional four residues at the N terminus ($M_r = 16,046$), was confirmed by SDS–PAGE and mass spectroscopy. NMR samples at 1–1.5 mM concentration were prepared in 50 mM sodium phosphate buffer (pH 6.8), 150 mM NaCl, 0.2 mM dithiothreitol in H₂O or ²H₂O.

NMR spectroscopy and structure determination

NMR spectra were acquired at 22 °C on Bruker DRX500, DRX600, DRX700 or DRX900 spectrometers equipped with triple resonance or cryogenic probes (500 and 600 MHz). Spectra were processed with NMRPipe¹⁹ and analysed using NMRVIEW²⁰. The ¹H, ¹³C and ¹⁵N chemical shifts were assigned by standard methods²¹. Distance restraints were derived from ¹⁵N- or ¹³C-resolved three-dimensional NOESY. H^N–N residual dipolar couplings were measured using a spin-state-selective ¹H–¹⁵N correlation experiment. Restraints for the backbone angles ϕ and ψ were derived from ³J(H^N,N) couplings and TALOS²². Slow-exchanging amide protons were identified from ¹H–¹⁵N correlation experiments after dissolving of lyophilized protein in ²H₂O. The experimentally determined distance and dihedral and dipolar coupling restraints (Table 1) were applied in a simulated annealing protocol using ARIA²³ and CNS²⁴. The final ensemble of NMR structures was refined in a shell of water molecules²⁵. Structural quality was analysed using PROCHECK²⁶. Ribbon and surface representations were prepared with MOLMOL²⁷ and GRASP²⁸, respectively.

Mutagenesis and ultraviolet cross-linking assays

For ultraviolet cross-linking assays, the PAZ domain of Ago2 was expressed as a glutathione S-transferase (GST) fusion using pGEXCS (pGEXCS–Ago2–PAZ). Ago2–PAZ mutants were generated in this vector using an oligonucleotide-directed *in vitro* mutagenesis system from Stratagene (quick-change site-directed mutagenesis). The recombinant proteins were expressed in *E. coli* BL21(DE3) pLysS as described²⁹. The final concentration of recombinant proteins used in each binding reaction is indicated in the figure legends. Ultraviolet cross-linking reactions were carried out in PBS with a 5' end-labelled siRNA (siRNA–2s) (30,000 counts per min reaction). Final sample volumes were 10 μ l. After 10 min incubation at 30 °C, samples were exposed to ultraviolet light (300 mJ). Reactions were stopped with 10 μ l of 2 $\times$ SDS loading buffer, heated for 5 min at 100 °C, and resolved on a 12% SDS–PAGE (37.5:1, acryl:bisacryl ratio). Gels were stained with Coomassie blue to control for equivalent loading, fixed, dried and visualized by autoradiography.

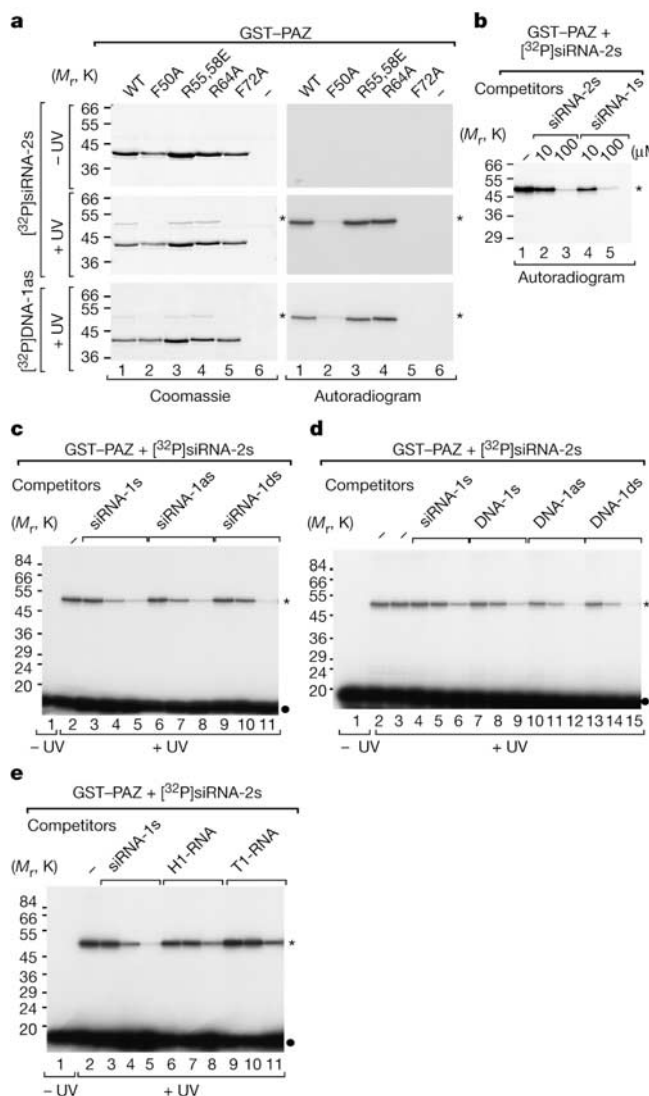


Figure 3 Nucleic-acid binding activity of the Ago2 PAZ domain. **a**, Comparative analysis of PAZ domain mutants after ultraviolet (UV) cross-linking to [³²P]-end-labelled oligonucleotides. Proteins indicated above the lanes were incubated with siRNA–2s (top) or DNA–1as (bottom) and irradiated with ultraviolet light. Samples were analysed by SDS–PAGE. The cross-linked protein–RNA complexes (asterisks) were visualized by Coomassie staining (left panels) and autoradiography (right panels). The concentration of recombinant proteins and oligonucleotides were 5 and 100 μ M, respectively. **b–e**, Cross-linking experiments using different nucleic acid ligands. The PAZ domain of Ago2 was incubated with [³²P]-end-labelled siRNA–2s (<5 nM) in the presence of increasing amounts of the competitors indicated above the lanes. The concentration of the recombinant PAZ domain in the reaction mixtures was 5 μ M. The competitors were tested at 10 and 100 μ M (**b**), 10, 50 and 100 μ M (**c, d**) and at 20, 100 and 200 μ M (**e**). The positions of the free siRNA–2s (black dot) and of the siRNA–2s–protein adducts (asterisk) is shown on the right. The sequences of the oligonucleotides are listed in Supplementary Table S1.

Received 10 August; accepted 15 October 2003; doi:10.1038/nature02123.
Published online 16 November 2003.

- Denli, A. M. & Hannon, G. J. RNAi: an ever-growing puzzle. *Trends Biochem. Sci.* **28**, 196–201 (2003).
- Hutvagner, G. & Zamore, P. D. RNAi: nature abhors a double-strand. *Curr. Opin. Genet. Dev.* **12**, 225–232 (2002).
- Hammond, S. M., Bernstein, E., Beach, D. & Hannon, G. J. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* **404**, 293–296 (2000).
- Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**, 25–33 (2000).
- Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**, 363–366 (2001).
- Elbashir, S. M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15**, 188–200 (2001).
- Nykanen, A., Haley, B. & Zamore, P. D. ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell* **107**, 309–321 (2001).
- Hammond, S. M., Boettcher, S., Caudy, A. A., Kobayashi, R. & Hannon, G. J. Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science* **293**, 1146–1150 (2001).
- Martinez, J., Patkaniowska, A., Urlaub, H., Lührmann, R. & Tuschl, T. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell* **110**, 563–574 (2002).
- Cerutti, L., Mian, N. & Bateman, A. Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain. *Trends Biochem. Sci.* **25**, 481–482 (2000).
- Kataoka, Y., Takeichi, M. & Uemura, T. Developmental roles and molecular characterization of a *Drosophila* homologue of *Arabidopsis* Argonaute1, the founder of a novel gene superfamily. *Genes Cells* **6**, 313–325 (2001).
- Carnell, M. A., Xuan, Z., Zhang, M. Q. & Hannon, G. J. The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes Dev.* **16**, 2733–2742 (2002).
- Tabara, H. et al. The rde-1 gene, RNA interference, and transposon silencing in *C. elegans*. *Cell* **99**, 123–132 (1999).
- Theobald, D. L., Mitton-Fry, R. M. & Wuttke, D. S. Nucleic acid recognition by OB-fold proteins. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 115–133 (2003).
- Volpe, T. A. et al. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**, 1833–1837 (2002).
- Yan, K. S. et al. Structure and conserved RNA binding of the PAZ domain. *Nature* advance online publication, 16 November 2003 (doi:10.1038/nature02129).
- Zhang, H., Kolb, F. A., Brondani, V., Billy, E. & Filipowicz, W. Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP. *EMBO J.* **21**, 5875–5885 (2002).
- Elbashir, S. M., Martinez, J., Patkaniowska, A., Lendeckel, W. & Tuschl, T. Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO J.* **20**, 6877–6888 (2001).
- Delaglio, F. et al. NMRPipe: a multidimensional spectral processing system based on UNIX Pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
- Johnson, B. A. & Blevins, R. A. NMRView: A computer program for the visualization and analysis of NMR data. *J. Biomol. NMR* **4**, 603–614 (1994).
- Sattler, M., Schleucher, J. & Griesinger, C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. NMR Spectrosc.* **34**, 93–158 (1999).
- Cornilescu, G., Delaglio, F. & Bax, A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13**, 289–302 (1999).
- Linge, J. P., O'Donoghue, S. I. & Nilges, M. Automated assignment of ambiguous nuclear Overhauser effects with ARIA. *Methods Enzymol.* **339**, 71–90 (2001).
- Brünger, A. T. et al. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Linge, J. P., Williams, M. A., Spronk, C. A., Bonvin, A. M. & Nilges, M. Refinement of protein structures in explicit solvent. *Proteins* **50**, 496–506 (2003).
- Laskowski, R. A., Rullmann, J. A., MacArthur, M. W., Kaptein, R. & Thornton, J. M. AQUA ROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8**, 477–486 (1996).
- Koradi, R., Billeter, M. & Wüthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55 (1996).
- Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Protein Struct. Funct. Genet.* **11**, 281–296 (1991).
- Grüter, P. et al. TAP, the human homolog of Mex67p, mediates CTE-dependent RNA export from the nucleus. *Mol. Cell* **1**, 649–659 (1998).
- Bogden, C. E., Fass, D., Bergman, N., Nichols, M. D. & Berger, J. M. The structural basis for terminator recognition by the Rho transcription termination factor. *Mol. Cell* **3**, 487–493 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. Ladurner, E. Conti and R. Russell for discussions; G. Stier and S. Bäckström for the pETM160 vector; M. Rode for technical support; F. Ciccarelli and P. Bork for help with sequence alignments; and the NMR centre in Frankfurt, Germany for NMR measurement time. This study was supported by the European Molecular Biology Organization (EMBO), the German Research Foundation (DFG), and the Human Frontier Science Program Organization.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.I. (izauralde@embl.de) or M.S. (sattler@embl.de). The coordinates and NMR data of the Ago2 PAZ domain have been deposited in the Protein Data Bank under accession number 1upo.

Structure and conserved RNA binding of the PAZ domain

Kelley S. Yan, Sherry Yan, Amjad Farooq, Arnold Han, Lei Zeng & Ming-Ming Zhou

Structural Biology Program, Department of Physiology and Biophysics, Mount Sinai School of Medicine, New York University, One Gustave L. Levy Place, New York, New York 10029-6574, USA

The discovery of RNA-mediated gene-silencing pathways, including RNA interference^{1–3}, highlights a fundamental role of short RNAs in eukaryotic gene regulation^{4–10} and antiviral defence^{11,12}. Members of the Dicer and Argonaute protein families are essential components of these RNA-silencing pathways^{13–19}. Notably, these two families possess an evolutionarily conserved PAZ (Piwi/Argonaute/Zwille) domain whose biochemical function is unknown. Here we report the nuclear magnetic resonance solution structure of the PAZ domain from *Drosophila melanogaster* Argonaute 1 (Ago1). The structure consists of a left-handed, six-stranded β -barrel capped at one end by two α -helices and wrapped on one side by a distinctive appendage, which comprises a long β -hairpin and a short α -helix. Using structural and biochemical analyses, we demonstrate that the PAZ domain binds a 5-nucleotide RNA with 1:1 stoichiometry. We map the RNA-binding surface to the open face of the β -barrel, which contains amino acids conserved within the PAZ domain family, and we define the 5'-to-3' orientation of single-stranded RNA bound within that site. Furthermore, we show that PAZ domains from different human Argonaute proteins also bind RNA, establishing a conserved function for this domain.

The three-dimensional structure of the minimal PAZ domain of *D. melanogaster* Ago1 (residues 298–427) is well defined by nuclear magnetic resonance (NMR) (Fig. 1a; see also Supplementary Table 1). Its compact architecture is built on a left-handed, six-stranded β -barrel core consisting of amino acid sequences conserved within the PAZ domain family²⁰ (Fig. 1b, c; see also Supplementary Fig. 1). Helices α 1 and α 2 pack together to close off one end of the barrel by forming interactions between Val 305, Met 309 and Phe 333, and the hydrophobic core of the barrel. A long β -hairpin (β 5– β 6) and a short helix α 3 form a structural appendage that is inserted between β 4 and β 7, and wraps around the rim of the β -barrel along β 8 (Fig. 1b, c). Although this appendage is an integral part of the PAZ domain structure, residues in β 5 and β 6 exhibit high dynamics (Supplementary Fig. 2). These structural elements are tied together by the amino- and carboxy-terminal regions that form a small anti-parallel β -sheet (β 1 and β 10), emphasizing the modular nature of this domain. Notably, although the conserved C-terminal residues 430–464 (Supplementary Fig. 1) are predicted to be a long α -helix, our NOE (nuclear Overhauser effects) analysis shows that they are unstructured (Supplementary Fig. 2). Truncation of this C-terminal segment did not affect the structure, confirming that it is not essential for the isolated PAZ domain.

Sequence conservation implies that the β -barrel also serves as the structural core for other Argonaute family PAZ domains, and the appendage may also be present with a β -hairpin of variable length (Supplementary Fig. 1). Significant variations exist in the loop sequences of different PAZ domains, possibly reflecting differences in cellular function. Dicer PAZ domains exhibit major sequence deviations from the Argonaute PAZ domains in the β 7– β 8 loop, although they probably still adopt the basic β -barrel and α 3 structure. The *D. melanogaster* Ago1 PAZ domain has a highly positive electrostatic potential (Fig. 1d), but most of its Arg and Lys residues are not conserved in other PAZ domains. Highly conserved residues including Tyr 354, Pro 363 and Tyr 386 are clustered at the

entrance of the β -barrel (Figs 1c and 3a; see also Supplementary Fig. 1), suggesting a possible functional site. Owing to the high level of structural conservation, we define the β -barrel and its associated appendage as the PAZ fold.

Despite its suggested function in protein–protein interaction¹⁷, the β -barrel core of the PAZ fold topologically resembles that of the RNA-binding domain of Sm proteins²¹ (Supplementary Fig. 3). We explored the possibility of nucleic-acid binding using NMR titration, and found that the Ago1 PAZ domain binds single-stranded (ss)RNA oligonucleotides with high preference over ssDNA counterparts (Fig. 2a). As demonstrated in two-dimensional ¹H-¹⁵N-HSQC (heteronuclear single quantum coherence) spectra (Fig. 2b), the protein forms a 1:1 stoichiometric complex with a 5'-pUGACA ssRNA (dissociation constant (K_d) approximately 1 μ M), a consensus sequence found at the 5' ends of microRNAs (miRNAs)²². The PAZ domain binds this ssRNA marginally better than the corresponding double-stranded (ds)RNA, and this interaction does not require the unstructured C-terminal segment (data not shown). Moreover, the RNA binding appears independent of sequence, as the protein binds equally well to different RNAs including a 5'-pUGAGG derived from the 5' end of *let-7* miRNA²³. The length of RNA does affect binding. Addition of ssRNA longer than 9–11 bases, or of prototypical 21-nt siRNA duplexes with 5' phosphates and 2-nt 3' overhangs, results in severe line broadening of NMR signals for the vast majority of the PAZ domain residues that is reversible with RNase treatment (data not

shown). This line-broadening phenomenon may be explained by: multiple PAZ domain molecules binding to a single RNA molecule, forming a complex analogous to 'protein beads on an RNA string'; or by a single PAZ domain engaging in different modes of interactions with a single RNA molecule (that is, 'sliding' through the RNA sequence), resulting in coexistence of different complex species. Both possibilities are consistent with the sequence-independent RNA binding of the PAZ domain.

Short regulatory RNAs have characteristic 5' phosphate and 3' hydroxyl groups, and 5' phosphorylation is required for incorporation into the RNA-induced silencing complex (RISC) in RNA interference^{24–27}. Indeed, comparison of the binding of the Ago1 PAZ domain to 5'-pUGACA and 5'-UGACA revealed that although the two ssRNAs use the same mode of interaction, as assessed by the similar pattern of chemical shift perturbations, the 5' phosphorylated RNA results in additional resonance changes and reduced line broadening for a subset of protein residues (Fig. 2c), possibly indicative of higher affinity binding. Modification of the 5' phosphate group by an amino-methylene (3-carbon) linkage²⁷ reduces the binding affinity (K_d approximately 2 μ M). These observations suggest that 5' phosphorylation may contribute to the PAZ domain–RNA interaction. Preliminary data using a 5' phosphorylated ssRNA with a 3'-biotin modification showed that immobilized protein failed to pull down purified Ago2 PAZ domain (data not shown). By NMR titration, this substrate had a reduced binding affinity (K_d greater than 20 μ M), suggesting that the 3' end may also

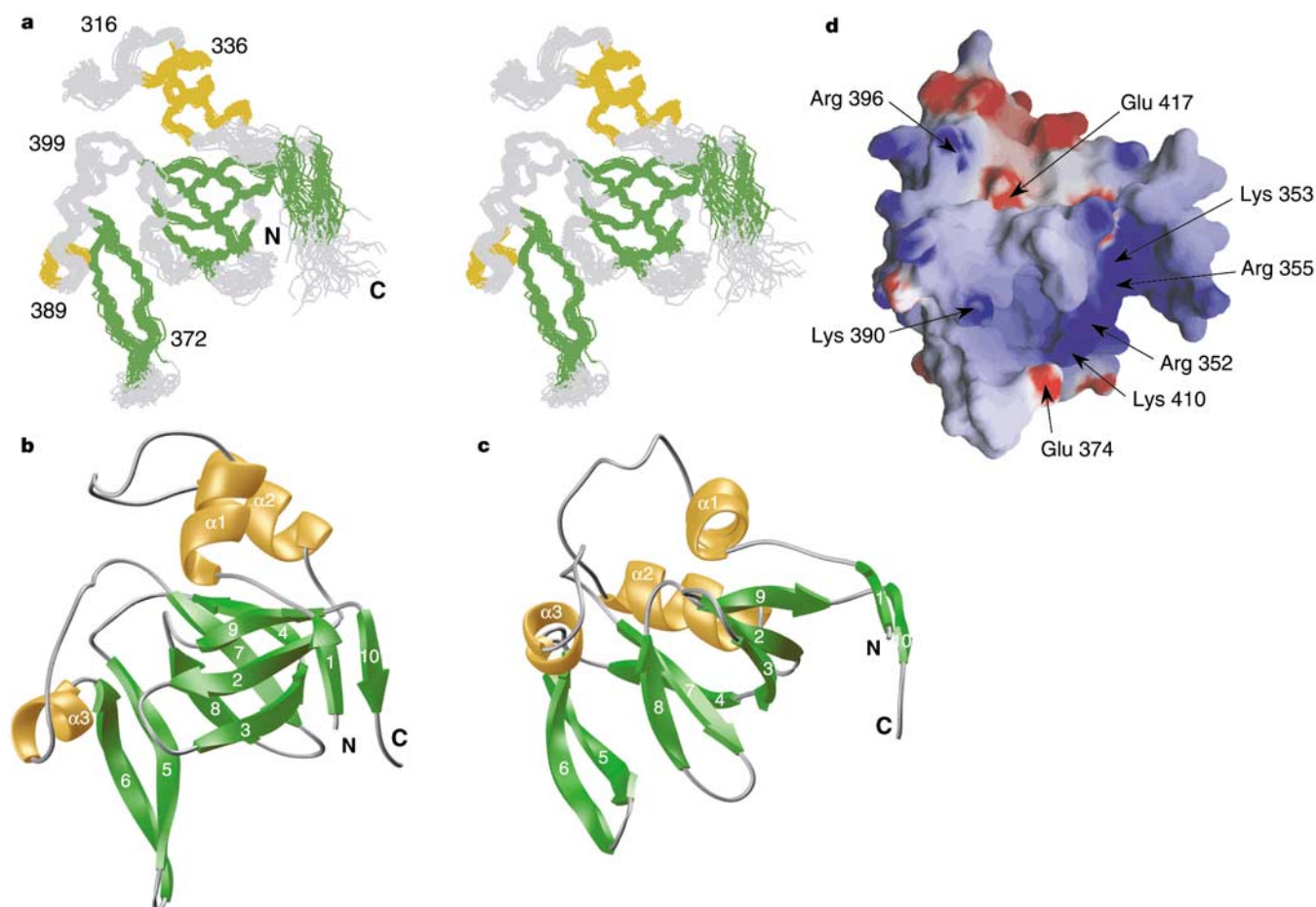


Figure 1 Three-dimensional structure of the *D. melanogaster* Ago1 PAZ domain. **a**, Stereoview of the backbone atoms of 25 superimposed NMR-derived structures of the PAZ domain (residues 298–430). The C-terminal unstructured residues (431–464) are omitted for clarity. **b**, Ribbons depiction (side view) of a representative NMR structure. The

orientation of **b** is similar to that in **a**. **c**, Top view of the PAZ domain. **d**, The surface electrostatic potential (blue, positive; red, negative) of the PAZ domain calculated in GRASP, displayed in the same orientation as **c**.

play a role in PAZ domain interaction.

We further characterized the Ago1 PAZ domain–RNA interaction with an RNA electrophoretic mobility shift assay (EMSA). However, possibly due to Ago1's high isoelectric point of 9.5 and the nature of RNA interactions, the ribonucleoprotein complex was not directly detected by EMSA. As shown in the NMR study, this PAZ domain can form a 1:1 stoichiometric complex with short RNA (about 5–7 nucleotides), or a 'protein beads on an RNA string' complex with longer RNA. In both cases, the complex is too positively charged at pH 8.3 to migrate towards the positive electrode in an EMSA. This reasoning seems consistent with the observation of a stable protein–RNA complex in an EMSA with human Piwi-like 1 PAZ domain (pI of 7.0; see below). Alternatively, this lack of detection of the *D. melanogaster* Ago1 PAZ domain–RNA complex might be due to the kinetic instability of the complex. However, we observed that the free ^{32}P -labelled RNA band decreases as a function of increasing Ago1 PAZ domain concentration (Fig. 2d), indicating formation of a protein–RNA complex. This RNA binding could be effectively competed away by unlabelled RNA of identical sequence, but to a lesser extent by DNA (Fig. 2d). These data confirm that the *D. melanogaster* Ago1 PAZ domain is able to interact with RNA preferentially over DNA.

The PAZ domain residues that undergo chemical shift changes upon RNA binding are localized at the open end of the β -barrel and the appendage, which together form an elongated binding groove (Fig. 3a, b). This cleft contains many of the residues highly conserved among PAZ domains. Because the PAZ domain binds 5' phosphorylated and 5' hydroxyl ssRNA in a similar manner, differential chemical shift perturbation analysis enabled us to define the 5'-to-3' orientation of the ssRNA bound within this groove (Fig. 2c). Out of a small number of residues, Gln 407 and Arg 355 showed the most profound differences in chemical shift perturbations, suggesting that the 5' end of a ssRNA molecule binds to a site in the region of the β 3– β 4 loop and the end of β 7. The 3' end binding site possibly involves an aromatic cluster at α 3 containing the conserved Tyr 386, Phe 387 and Tyr 391 residues. End to end, the RNA-binding groove spans approximately 35–40 Å in length, corresponding to the length of a 5–7-nucleotide RNA that forms a stable 1:1 complex with the PAZ domain.

Using NMR, we probed the molecular nature of PAZ domain–RNA interactions with UMP (uridine 5'-monophosphate). Despite its low affinity (K_d approximately 2 mM), UMP is bound in the predicted RNA-binding groove, as revealed by the structure of the PAZ domain–UMP complex (Fig. 3c, d). The 2' and 3' hydroxyl groups are positioned to form hydrogen bonds with the backbone atoms of His 411 and Thr 412 of β 8 (Fig. 3d). The aromatic uracil interacts with the terminal nitrogen of Lys 410, and the 5' phosphate of UMP is possibly coordinated by the nitrogen atoms of Arg 352. The intermolecular NOEs to the ribose are greater in number (11 compared with 2) and intensity than those seen to the base, indicating that recognition of UMP involves mostly the sugar moiety rather than the base. These observations agree with our finding of the PAZ domain preference of RNA over DNA and the apparent lack of sequence dependence. As the intermolecular NOEs observed for ssRNA also involve the same set of residues conserved in sequence, the PAZ domain–UMP complex structure provides valuable insights into RNA recognition.

To investigate the molecular determinants of RNA binding by the Ago1 PAZ domain, we performed site-directed mutagenesis on 15 amino acid residues that are localized to the RNA-binding groove and conserved in the PAZ domain family. The effect of these mutations on RNA binding was assessed by the ability of the mutants to compete with wild-type PAZ domain for binding to ssRNA (Fig. 3e; see also Supplementary Fig. 4). Individual mutation of Tyr 354, Met 366, Lys 390, Arg 392, Arg 396 or Glu 417 to Ala did not cause major reductions in RNA binding. However, point mutation of His 346, Arg 352, Phe 369, Lys 385, Tyr 386, Phe 387,

Tyr 391 or Lys 410 to Ala resulted in a greater than 50% loss of RNA binding as compared with the wild type. Notably, R355A lost nearly 90% of the RNA-binding activity. The functionally important residues as defined by mutagenesis are clustered in three regions of the PAZ fold (Fig. 3a, e): (1) the β 3– β 4 loop (Arg 355), a binding site for the 5' end of ssRNA; (2) the rim of the β -barrel orifice formed by the β 2– β 3 and β 7– β 8 loops (His 346, Arg 352 and Lys 410), where UMP binds; and (3) the solvent-exposed basic and

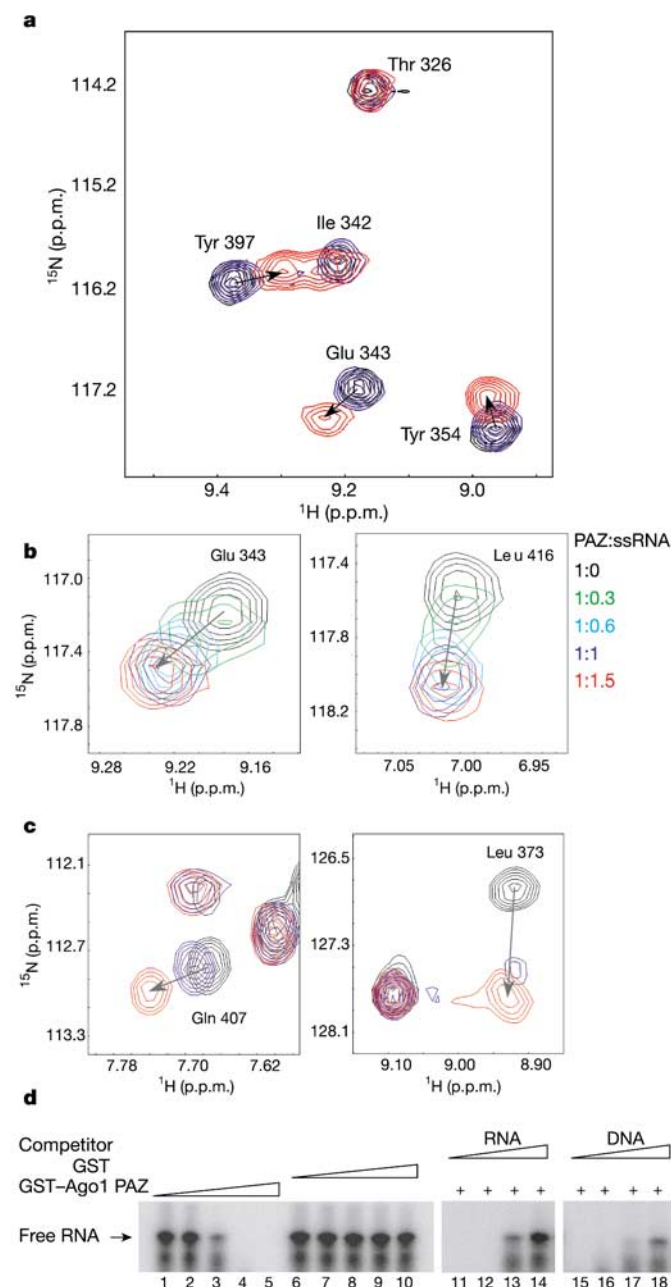


Figure 2 RNA binding of the *D. melanogaster* Ago1 PAZ domain. **a**, Superimposition of ^1H - ^{15}N -HSQC spectra of the PAZ domain (approximately 0.3 mM) in free form (black), with 5' phosphorylated ssDNA (5'-pTGACA; PAZ:ligand molar ratio 1:1.5) (blue) or 5' phosphorylated ssRNA (5'-pUGACA; 1:1.5) (red). p.p.m., parts per million. **b**, Stoichiometry of PAZ domain–ssRNA binding. The amide signals of Glu 343 and Leu 416 in the superimposed NMR spectra are colour-coded according to PAZ:RNA molar ratio. **c**, Binding of 5' hydroxyl (blue) compared with 5' phosphorylated (red) ssRNA (UGACA). The PAZ:ssRNA molar ratio was 1:1.5. **d**, RNA electrophoretic mobility shift assay, assessing the Ago1 PAZ domain interaction with RNA and DNA.

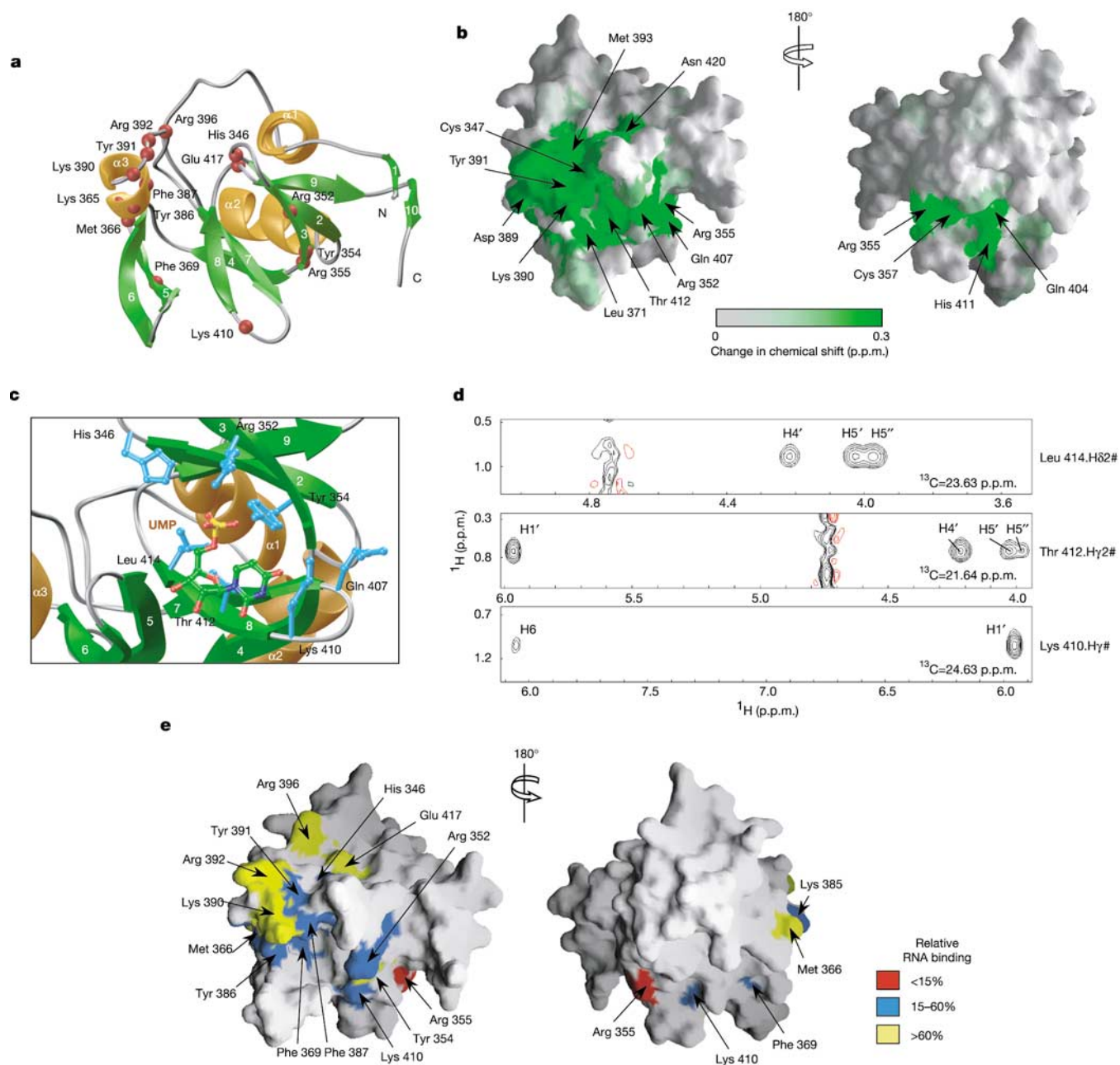


Figure 3 Mapping the RNA-binding site of the *D. melanogaster* Ago1 PAZ domain. **a**, Structure of the PAZ domain depicting residues (red spheres) conserved in the PAZ domain family. **b**, Molecular surface of the protein highlighting residues that exhibit backbone ^1H and ^{15}N chemical shift perturbations on binding to ssRNA (5'-pUGACA; PAZ:RNA 1:1.5). The orientation of the left panel is same as **a**. **c**, UMP binding site of the PAZ domain–UMP complex structure. **d**, Intermolecular NOEs between the PAZ domain and UMP observed in a three-dimensional ^{13}C -edited, $^{13}\text{C}/^{15}\text{N}$ -filtered NOESY spectrum.

H82#, Hγ2# and Hγ# represent specific side-chain hydrogen atoms in amino acid residues. **e**, Mutational analysis of the PAZ domain. Effects of individual mutations on RNA binding were assessed by an inhibition experiment described in the Methods and Supplementary Fig. 2, and colour-coded according to their relative RNA-binding activities. Results were normalized to 100% RNA-binding activity with the wild-type protein.

aromatic cluster at $\alpha 3$ (Lys 385, Tyr 386, Phe 387 and Tyr 391). The large number and extended distribution of possible contact points between the PAZ domain and RNA may explain the lack of a null point mutant, and confirm our findings on the location and geometry of the RNA-binding site in the PAZ fold.

To test whether RNA binding is a general function of the PAZ domain, we prepared PAZ domains from different subgroups of proteins within the large PAZ domain family. As predicted, the PAZ domains from human Ago1 (eIF2C1), Ago2 (eIF2C2) and Piwi-like 1 bind to ssRNA (Fig. 4a–c). Similar to *D. melanogaster* Ago1, the human PAZ domains bind RNA preferentially over DNA, and form

a 1:1 stoichiometric complex with 5-nucleotide ssRNA. Moreover, the Piwi-like 1 PAZ domain–RNA complex was directly detected by EMSA (Fig. 4d). The band corresponding to the complex increases as a function of protein concentration, whereas the free RNA signal decreases accordingly. The Piwi-like 1 PAZ domain–RNA complex was effectively competed away with unlabelled RNA, but not as well with the corresponding DNA (Fig. 4d). From these results, we conclude that RNA binding is a conserved function of the PAZ domain.

The structure and RNA binding of the *D. melanogaster* Ago2 PAZ domain reported in the accompanying paper²⁸ are consistent with

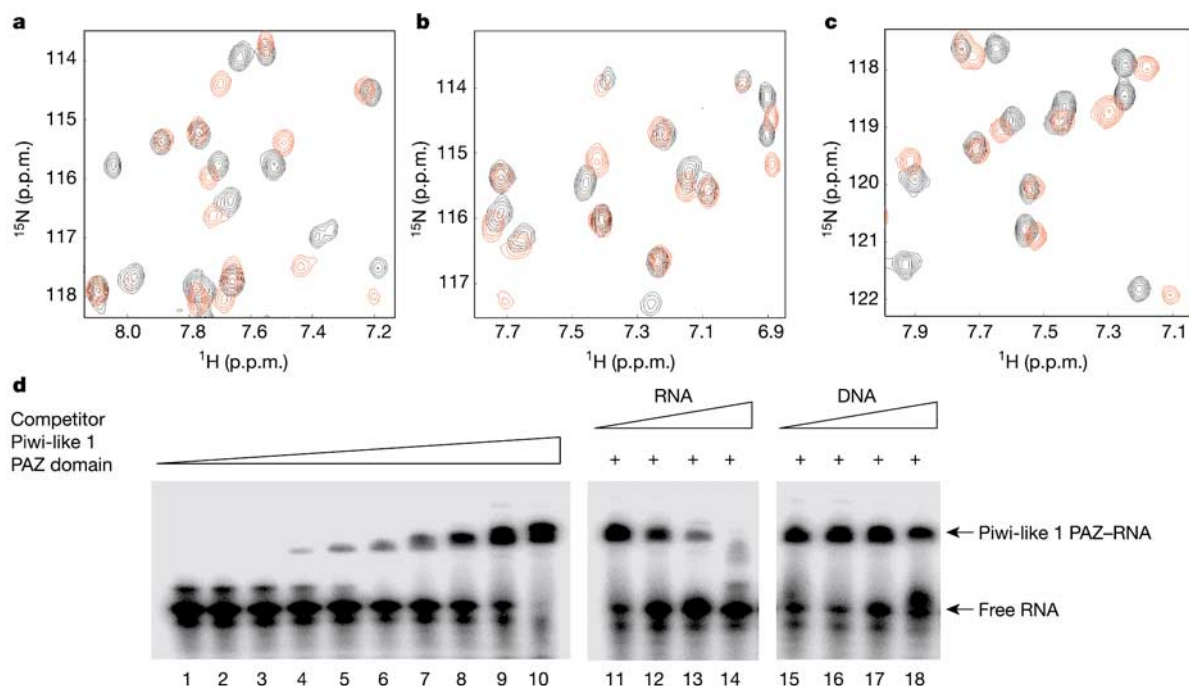


Figure 4 Conserved RNA binding of the PAZ domain. **a–c**, Superimposition of ^1H - ^{15}N -HSQC spectra of the PAZ domains from human Ago1 (**a**), Ago2 (**b**) and Piwi-like 1 (**c**) in free form (black) and with a 5'-pUGAGG ssRNA (red, PAZ:RNA 1:1).

d, RNA electrophoretic mobility shift assay assessing RNA binding of human Piwi-like 1 PAZ domain.

the predictions from our study that the PAZ fold and RNA binding are conserved in the PAZ domain family. The molecular nature and specificity of DNA binding shown by the *D. melanogaster* Ago2 PAZ domain, which is in contrast to the preference of RNA over DNA binding by the Ago1 PAZ domain demonstrated in our study, however, remains to be evaluated for its possible biological importance.

Our structure, mutagenesis and RNA mobility shift data of the PAZ domain define this module as an RNA-binding domain, thereby establishing the molecular function of this family. The PAZ domain uses a six-stranded β -barrel, similar to Sm proteins²¹ but with an additional appendage, to bind both single- and double-stranded RNA. Although RNA binding by the PAZ domain *in vitro* does not appear to be sequence-dependent, *in vivo* specificity may be conferred within the context of the full-length protein or through accessory proteins within RISC. The unique features of short regulatory RNAs, such as their length of approximately 22 nucleotides and 5' phosphate and 3' hydroxyl ends may contribute to specificity^{13,24–27}. Our data show that these 5' and 3' ends may enhance RNA binding by the PAZ domain, and that the binding groove accommodates RNA of 5–7 nucleotides in length. These findings may correlate with the 5' phosphorylation requirement of siRNAs or miRNAs for their incorporation into RISC^{24,27}, as well as the observed length of the 5' end sequence conservation in miRNAs^{22,29}. □

Methods

Sample preparation

The PAZ domain of *Drosophila* Argonaute 1 (residues 297–464 or 297–433) was cloned into a pET19b vector and expressed in *Escherichia coli* BL21(DE3) cells as an N-terminal His₁₀-tagged protein. Uniformly ^{15}N - and $^{15}\text{N}/^{13}\text{C}$ -labelled proteins were prepared by growing bacteria in minimal medium with $^{15}\text{NH}_4\text{Cl}$ and/or $^{13}\text{C}_6$ -glucose as the sole nitrogen and carbon sources. Deuterated protein was generated by cell growth in 95% $^2\text{H}_2\text{O}$. The PAZ domain was purified by nickel-NTA affinity chromatography. After cleavage of the His₁₀ tag, the protein was further purified by ion-exchange and size-exclusion chromatography. Protein NMR samples (approximately 0.5 mM) were prepared in 100 mM phosphate buffer of pH 6.5, 150 mM NaCl and 5 mM dithiothreitol (DTT)-*d*₁₀

in $\text{H}_2\text{O}/^2\text{H}_2\text{O}$ (9/1) or $^2\text{H}_2\text{O}$. PAZ domains from the human proteins Ago1 (residues 218–356), Ago2 (residues 220–358) and Piwi-like 1 (residues 604–910) were cloned into the pET19b expression vector. These proteins were expressed and purified using a procedure similar to that of the *D. melanogaster* Ago1 PAZ domain. Glutathione S-transferase (GST)–Ago1 PAZ domain fusion proteins were expressed in bacteria using the pGEX-6P-1 vector, and purified by affinity chromatography.

NMR spectroscopy

The Ago1 PAZ domain containing residues 297–464 was used for structure determination. All NMR spectra were acquired at 25 °C on a Bruker 500, 600, or 800 MHz spectrometer. The backbone ^1H , ^{13}C and ^{15}N resonances were assigned using standard three-dimensional triple-resonance HNCA, HN(CO)CA, HN(CA)CB and HN(COCA)CB experiments. The side-chain atoms were assigned from three-dimensional HCCH-TOCSY, HCCH-COSY and (H)C(CO)NH-TOCSY data. The NOE-derived distance restraints were obtained from ^{15}N - or ^{13}C -edited three-dimensional NOESY spectra. The $^3J(\text{H}^{\text{N}}, \text{H}^{\alpha})$ coupling constants measured from three-dimensional HNHA data were used to determine ϕ -angle restraints. Slowly exchanging amide protons were identified from a series of two-dimensional ^{15}N -HSQC spectra recorded after $\text{H}_2\text{O}/^2\text{H}_2\text{O}$ buffer exchange. The intermolecular NOEs used in defining the structure of the Ago1 PAZ domain-UMP complex were detected in ^{13}C -edited (F_1), $^{13}\text{C}/^{15}\text{N}$ -filtered (F_3) three-dimensional NOESY spectra.

Structure calculations

Structures of the PAZ domain were calculated with a distance geometry simulated annealing protocol with X-PLOR³⁰. Initial protein structure calculations were performed with manually assigned NOE-derived distance restraints. Hydrogen-bond distance restraints, generated from the H/D exchange data, were added at a later stage of structure calculations for residues with characteristic NOE patterns. The converged structures were used for the iterative automated NOE assignment by ARIA for refinement³¹. Structure quality was assessed with Procheck-NMR. The structure of the Ago1 PAZ domain-UMP complex was determined using the free-form structure in addition to 13 intermolecular NOE-derived distance restraints and four hydrogen-bond distance restraints.

RNA-binding assays

RNA and DNA oligonucleotide titration was performed by recording a series of two-dimensional ^1H - ^{15}N -HSQC spectra with a uniformly ^{15}N -labelled PAZ domain (0.3 mM) in the presence of increasing amounts of RNA or DNA ranging from 0 to 0.9 mM. The protein sample and the RNA/DNA stock solutions were all prepared in the same buffer containing 100 mM sodium phosphate, 150 mM NaCl and 5 mM DTT-*d*₁₀ at pH 6.5. Electrophoretic mobility shift assays were performed using a 26-nucleotide ssRNA (5'-AUUUUG-UUGUCGAAAUUGUACAUAA-3') labelled at the 5' end using T4 polynucleotide kinase and [γ - ^{32}P]ATP. The protein/RNA samples were run on native TBE polyacrylamide gels and visualized by autoradiography. The concentration of the 5' ^{32}P -labelled ssRNA was 2 μM in each EMSA experiment with a PAZ domain protein of

0–15 μ M. For the competition EMSA, the corresponding unlabelled 26-nucleotide RNA or DNA of 0–30 μ M was used in addition to the 32 P-labelled RNA and PAZ domain protein.

Site-directed mutagenesis

Expression constructs for mutant proteins were generated using the QuikChange site-directed mutagenesis kit (Stratagene). The presence of appropriate mutations was confirmed by DNA sequencing. Mutant proteins were purified by affinity chromatography using nickel-NTA columns (Qiagen). Proper folding of each PAZ domain mutant was confirmed by NMR. Samples for the NMR competition assay were prepared by mixing 15 N-labelled wild-type PAZ domain (0.2 mM), ssRNA oligonucleotide (5'-pUGACA) and unlabelled mutant PAZ domain with a molar ratio of 1:1:5.

Received 4 August; accepted 16 October 2003; doi:10.1038/nature02129.

Published online 16 November 2003.

1. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
2. Dykxhoorn, D. M., Novina, C. D. & Sharp, P. A. Killing the messenger: short RNAs that silence gene expression. *Nature Rev. Mol. Cell Biol.* **4**, 457–467 (2003).
3. Hannon, G. J. RNA interference. *Nature* **418**, 244–251 (2002).
4. Hutvagner, G. *et al.* A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293**, 834–838 (2001).
5. Grishok, A. *et al.* Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* **106**, 23–34 (2001).
6. Lee, R. C., Feinbaum, R. L. & Ambros, V. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843–854 (1993).
7. Volpe, T. A. *et al.* Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**, 1833–1837 (2002).
8. Hall, I. M. *et al.* Establishment and maintenance of a heterochromatin domain. *Science* **297**, 2232–2237 (2002).
9. Mochizuki, K., Fine, N. A., Fujisawa, T. & Gorovsky, M. A. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. *Cell* **110**, 689–699 (2002).
10. Schramke, V. & Allshire, R. Hairpin RNAs and retrotransposon LTRs effect RNAi and chromatin-based gene silencing. *Science* **301**, 1069–1074 (2003).
11. Ratcliff, F., Harrison, B. D. & Baulcombe, D. C. A similarity between viral defense and gene silencing in plants. *Science* **276**, 1558–1560 (1997).
12. Waterhouse, P. M., Wang, M. B. & Lough, T. Gene silencing as an adaptive defence against viruses. *Nature* **411**, 834–842 (2001).
13. Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**, 25–33 (2000).
14. Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**, 363–366 (2001).
15. Knight, S. W. & Bass, B. L. A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in *Caenorhabditis elegans*. *Science* **293**, 2269–2271 (2001).
16. Hammond, S. M., Bernstein, E., Beach, D. & Hannon, G. J. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* **404**, 293–296 (2000).
17. Hammond, S. M., Boettcher, S., Caudy, A. A., Kobayashi, R. & Hannon, G. J. Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science* **293**, 1146–1150 (2001).
18. Tabara, H. *et al.* The rde-1 gene, RNA interference, and transposon silencing in *C. elegans*. *Cell* **99**, 123–132 (1999).
19. Williams, R. W. & Rubin, G. M. ARGONAUTE1 is required for efficient RNA interference in *Drosophila* embryos. *Proc. Natl Acad. Sci. USA* **99**, 6889–6894 (2002).
20. Cerutti, L., Mian, N. & Bateman, A. Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain. *Trends Biochem. Sci.* **25**, 481–482 (2000).
21. Kambach, C. *et al.* Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell* **96**, 375–387 (1999).
22. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294**, 858–862 (2001).
23. Reinhart, B. J. *et al.* The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403**, 901–906 (2000).
24. Nykanen, A., Haley, B. & Zamore, P. D. ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell* **107**, 309–321 (2001).
25. Elbashir, S. M., Martinez, J., Patkaniowska, A., Lendeckel, W. & Tuschl, T. Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO J.* **20**, 6877–6888 (2001).
26. Schwarz, D. S., Hutvagner, G., Haley, B. & Zamore, P. D. Evidence that siRNAs function as guides, not primers, in the *Drosophila* and human RNAi pathways. *Mol. Cell* **10**, 537–548 (2002).
27. Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R. & Tuschl, T. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell* **110**, 563–574 (2002).
28. Lingel, A., Simon, B., Izaurralde, E. & Sattler, M. Structure and nucleic-acid binding of the *Drosophila* Argonaute 2 PAZ domain. *Nature* advance online publication, 16 November 2003 (doi:10.1038/nature02123).
29. Dostie, J., Mourelatos, Z., Yang, M., Sharma, A. & Dreyfuss, G. Numerous microRNPs in neuronal cells containing novel microRNAs. *RNA* **9**, 180–186 (2003).
30. Brunger, A. T. *X-PLOR Version 3.1: A System for X-Ray Crystallography and NMR* (Yale Univ. Press, New Haven, Connecticut, 1993).

31. Nilges, M. & O'Donoghue, S. Ambiguous NOEs and automated NOE assignment. *Prog. NMR Spectrosc.* **32**, 107–139 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. W. Williams for providing a *D. melanogaster* Ago1 expressed sequenced tag clone, T. Tuschl for complementary DNAs encoding human Argonaute proteins, and T. A. Edwards for discussions. K.S.Y. is a recipient of a National Institutes of Health (NIH) predoctoral training grant fellowship. M.-M.Z. is supported by NIH grants, and is a member of the New York Structural Biology Center.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.-M.Z. (Ming-Ming.Zhou@mssm.edu). Coordinates for the NMR structure of the *D. melanogaster* Ago1 PAZ domain are deposited in the Protein Data Bank under ID code 1R4K.

addendum

Specific cytotoxic T cells eliminate cells producing neutralizing antibodies

Oliver Planz, Peter Seiler, Hans Hengartner & Rolf M. Zinkernagel

Nature **382**, 726–729 (1996).

The experiments in BALB/c mice using hybridoma technology reported in this Letter indicated that neutralizing-antibody-producing B cells and hybridomas could have been preferentially infected by lymphocytic choriomeningitis virus (LCMV), whereas B cells specific for the internal viral components were not. We concluded that during LCMV infection, B cells specific for the viral-surface antigens may be selectively and productively infected by LCMV through the membrane-anchored neutralizing-antibody receptor and are subsequently eliminated by virus-specific cytotoxic T lymphocytes. These experiments were performed in normal or CD8⁺ T-cell-depleted BALB/c H-2^d mice. In later experiments designed to generalize these findings, we studied the infection patterns of hybridomas using C57BL/6 mice, but the results were variable and not conclusive. We also failed to observe a preferential infection of B cells or of hybridomas generated from IgM or isotype-switchable transgenic mice with a high percentage of neutralizing-antibody-expressing B cells^{1,2}. In a third series of experiments with a recombinant virus that combines the envelope glycoprotein from vesicular stomatitis virus (VSV) with an LCMV replicon, we did not detect a preferential CD8⁺ T-cell-dependent depletion of the anti-VSV antibody response (D. Pinschewer *et al.*, unpublished observations). We have also not so far been able to reproduce our original findings in new experiments with LCMV and BALB/c (SPF) mice. We therefore wish to alert the community to these problems until further experiments have been performed. □

1. Seiler, P. *et al.* Enhanced virus clearance by early inducible lymphocytic choriomeningitis virus-neutralizing antibodies in immunoglobulin-transgenic mice. *J. Virol.* **72**, 2253–2258 (1998).
2. Hengartner, L. *et al.* Antiviral immune responses in gene-targeted mice expressing the immunoglobulin heavy chain of virus-neutralizing antibodies. *Proc. Natl Acad. Sci. USA* **100**, 12883–12888 (2003).

Genomics and proteomics

Variations on the theme of DNA and the proteins it makes.

Adenovirus Technology

Q-Biogene

www.adenovirus.com; www.qbiogene.com

Web-based reference site

Q-Biogene has introduced a website to provide researchers with up-to-date information regarding recombinant adenovirus and its use in research areas such as gene delivery, gene expression and protein production. It houses information about the history and development of recombinant adenovirus and serves as a reference site for breakthroughs in new adenovirus technology, applications and critical background information. Product details on the company's Q-mate inducible expression system for the expression, study and production of toxic proteins and its RCA-Free cell lines for recombinant adenovirus preparation in gene therapy are also featured. The site can be found at www.adenovirus.com.

A.C.E. sequencing buffer

Amresco

www.amresco-inc.com

In the pink

Amresco's A.C.E. sequencing buffer for DNA capillary sequencing is now tinted pink for easier identification. Researchers using it will be able to differentiate between the buffer and any other solution or water in the sequencer. The buffer is available in 1 and 10 sizes (the 10 version requires 1:10 dilution). The buffer can be used with the ABI 3100, 3700, and 3730 automated CE systems.

CEQ 8000 version 6.0

Beckman Coulter

www.beckmancoulter.com

Cut and run

This software program analyses DNA sequence data while automatically identifying and trimming off vector material, foreign sequences and poor quality bases. A quality values function helps researchers identify

and isolate the highest quality portions of reads to improve the accuracy and completeness of sequence assembly. Users set their own quality criteria; the program then applies these values in the trimming of undesirable data. It is suitable for use in genome sequencing assembly, polymorphism studies, expressed sequence tag (EST) analysis, and phylogenetic analysis.

Lipofectamine

Invitrogen

www.invitrogen.com

Nothing beastly about this reagent

Lipofectamine 2000 CD (chemically defined) transfection reagent from Invitrogen is 100% animal-origin free. It delivers high transfection efficiencies and protein expression levels while eliminating concerns about introducing unidentified animal components during transfection.

Genetic immunization

QED Bioscience

www.qedbio.com

For high-affinity specific antibodies

QED Bioscience offers customized antibody production through direct immunization with the customer's cDNA. *In vivo* expression of the antigen guarantees correct protein folding and post-translational modifications. The procedure saves researchers the time and expense of expressing and purifying proteins and provides control of the immune response to the target. Low-level, consistent production of target protein *in vivo* produces high-affinity, specific antibodies.

RiboPure

Ambion

www.ambion.com

Improve RNA isolation from yeast

The RiboPure yeast RNA isolation kit from Ambion provides high yields of RNA from yeast cells. The kit uses no enzymatic pre-treatment step to disrupt the yeast cells and thus does not alter expression profiles. It incorporates a DNA-free DNase treatment step and the resulting RNA is said to be exceptionally pure, containing little or no genomic DNA (<0.05%). RNA can be produced from *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Pichia pastoris* and *Candida albicans*.

Ultrafree-DA

Millipore

www.millipore.com

One-step DNA purification from agarose gels

Millipore's Ultrafree-DA centrifugal device is designed to maximize DNA recovery from agarose gel slices. Requiring a single-step, 10-



Band aid: Ultrafree-DA for DNA purification.

minute spin at 5,000g, the device delivers recoveries for DNA ranging in size from 100 to 10,000 base pairs. A band of interest is excised from an agarose gel and placed into the pre-assembled filter device. Centrifugal force compresses the agarose and extrudes the DNA from the gel pores. The gel matrix is retained by the membrane filter, while the DNA passes into a collection tube. The resulting purified DNA is fully functional and ready for sequencing, cloning or labelling.

Ettan kits and reagents

Amersham Biosciences

www.amershambiosciences.com

It's all in the preparation

Amersham Biosciences has expanded its Ettan proteomics product range to include sample preparation kits and reagents. Offerings include an albumin and IgG removal kit, 2-D fractionation kit, and basic chemicals and reagents. These kits and reagents remove contaminating compounds such as nucleic acids, carbohydrates and lipids, as well as abundant proteins that can mask the signals generated by important trace proteins. Applications include 2-D electrophoresis, mass spectrometry and western blotting.

Total & Poly A⁺ RNAs

BD Biosciences

www.bdbiosciences.com

Rare finds

BD Biosciences has expanded its line of Total & Poly A⁺ RNAs to include many samples from tissues that are rare or difficult to obtain. The RNAs include 22 human brain and 6 heart regions, as well as an extended collection of mouse and rat Total and PolyA⁺ RNAs. Each sample is prepared using the company's modified guanidium thiocyanate



Now they are six: the latest CEQ 8000 is now out.

method and enriched for mRNA transcripts with two rounds of oligo(dT)-cellulose purification. The RNAs are ready for any application, including quantitative PCR, cDNA synthesis, ribonuclease protection assays, library construction and northern blotting.

FluorChem 5500

Alpha Innotech www.alphainnotech.com

Economical solution for low-light applications

The FluorChem 5500 imaging system from Alpha Innotech features advanced noise reduction algorithms and a high signal-to-noise ratio, producing a wide 12-bit dynamic range. It offers resolution of more than 1.1 million pixels and can be used in fluorescence, chemiluminescence and visible light imaging applications such as 1-D gels, 2-D gels, films, membranes, plates and assays. The system is equipped with AlphaEase FC image analysis software. With real-time imaging, the 5500 enables users to focus and capture samples at a readout rate of 30 frames per second. Three separate binning



On the dark side: FluorChem 5500.

modes (1 1, 2 2 and 3 3) allow for desired image resolution with the necessary sensitivity.

Cellspin

Integra Biosciences

www.integra-biosciences.com

High-yield protein expression from cell suspensions

Cellspin from Integra Biosciences is a compact system that enables high-yield protein expression from mammalian or insect cells growing in suspension. Built to provide versatility for use in research- or laboratory-scale cell cultivation, Cellspin features stirring platforms that can accommodate four flasks across a range of sizes (100 ml, 250 ml, 500 ml and 1,000 ml). The system, comprised of a moisture-resistant flask deck and a removable control unit, allows cell cultivation to take place in both standard and CO₂ incubators. The system has been designed to be gentle on cells, with a rounded dual glass pendulum stirring mechanism. No heat is imparted to the cell suspension. Cellspin flasks provide a high surface-to-volume ratio to maximize oxygenation, while auto-centre coning ensures stability. The entire cell cultivation process can be pre-programmed through the control unit, with automatic adjustment of rotary velocity and angle, and the length and frequency of work cycles, allowing secure handling of sensitive cell lines.

GelPal

Genetix

www.genetix.com

Hits the spot

GelPal is a benchtop system for excision of protein spots from both backed and unbacked 1-D and 2-D polyacrylamide gels. It dispenses removed gel plugs into a 96-well microplate. Incorporating a hand-held excision pen with interchangeable heads of varying diameter and depth, the GelPal can be used to excise a wide range of spot diameters and gel thick-

nesses (0.5–2.0 mm). Using appropriate illumination, the device can be used to remove protein spots from Coomassie, silver or Sypro Ruby stained gels. Dispensing of gel plugs is controlled by a choice of hand or foot modules. To aid dispensing of gel plugs into the correct wells, the 96-well microplate is held on a microplate holder with illumination indent guides on either rows or columns, depending on user preference. The system is controlled and programmed using a simple keypad interface. The pick volume range (5–250 µl) and place volume range (5–250 µl) can be programmed, allowing users to define precisely the place volume they wish the gel plug to be stored in. Up to ten optimal gel excision programs can be stored. The software includes a purge function to ensure minimal cross-contamination of samples through efficient cleaning of the excision pin.

LUX primer sets

Invitrogen

www.invitrogen.com

Good housekeeping

Over 30 certified LUX primer sets for housekeeping genes are available from Invitrogen, who also manufacture LUX fluorogenic primers. The primer sets can be used for normalization of real-time PCR experiments. Each ready-to-use set is functionally validated and optimized for accurate quantification of a corresponding internal control gene. A wide range of genes, species and expression levels are available.

4700 Proteomics Discovery System

Applied Biosystems

www.europe.appliedbiosystems.com

Quick links to speed analysis

The 4700 Proteomics Discovery System provides researchers with information to help streamline protein identification, expression analysis and characterization. Experimentally identified proteins can be linked directly with the relevant biological information and annotative tools in the Celera DiscoverySystem online platform, including biological and molecular function, SNP sites, alternate splice forms and comprehensive genetic information. When combined with Results Dependent Analysis software, the system will direct further analysis of complex samples for additional discovery and characterization of low-abundance proteins when necessary. For example, researchers using ICAT reagents for protein expression studies can use the 4700 system to initially quantify the proteins present in a sample and then automatically resubmit only those proteins that are over- or under-expressed for additional MS/MS analysis and protein identification. The system fully supports LC-MALDI experiments.

These notes are compiled in the Nature office from information provided by the manufacturers.

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Netherlands/ Italy/ Spain/

Portugal/ Belgium:

Evelina Rubio Hakansson (4973)

Scandinavia: Sille Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Natureevents:

Paul Constant (4954)

Production Manager:

 Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager:

 Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Harakakatamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: rasami@naturejpn.com

naturejobs

Home-grown success

Gregor Mendel would have smiled upon the meeting of geneticists Jiri Bartek and Jiri Lukas. The pair first met in 1988 at the Cancer Research Institute, about 100 metres from the Brno monastery where Mendel famously cultivated his peas. Mendel would be even happier to hear that the two native Czech scientists are returning to their home country to claim an award bearing his name — the G. J. Mendel Honorary Medal for Merit in the Biological Sciences awarded by the Academy of Sciences of the Czech Republic.

The original meeting was pivotal for both their careers. They decided to join forces in Prague to establish a new department. But they soon found resources to be limited and scientific excellence “lacking”, says Bartek, so after a year they headed west for more training. Bartek went to the Imperial Cancer Research Fund (now Cancer Research UK) in London and Lukas joined the European Molecular Biology Laboratory in Heidelberg, Germany. “That was a turning point,” Bartek says. Both got access to facilities — and colleagues — they didn’t have back home.

Bartek and Lukas are now reunited at the Institute of Cancer Biology in Copenhagen, where they are working on the cell cycle and DNA damage and repair. When they return to Prague this week to receive their medals they will tell young scientists there that they now have more opportunities closer to home and that their skills are in demand, as interest in core disciplines in the West is slipping. The young Max Planck lab at Dresden now offers opportunities for scientists from the East (see *Nature* 413, 4–5; 2001) and is complemented by a host of growing facilities in Vienna (see *Naturejobs* 4–5; 11 April 2002). The two scientists will encourage aspiring researchers to make the same decision — leave home for better training in the best lab that will take you and then decide where to establish your own. But that advice holds true for young researchers everywhere.

Paul Smaglik
Naturejobs editor



Contents

CAREERS AND RECRUITMENT

The rise and rise of
nanotechnology p478

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Small world, big hopes

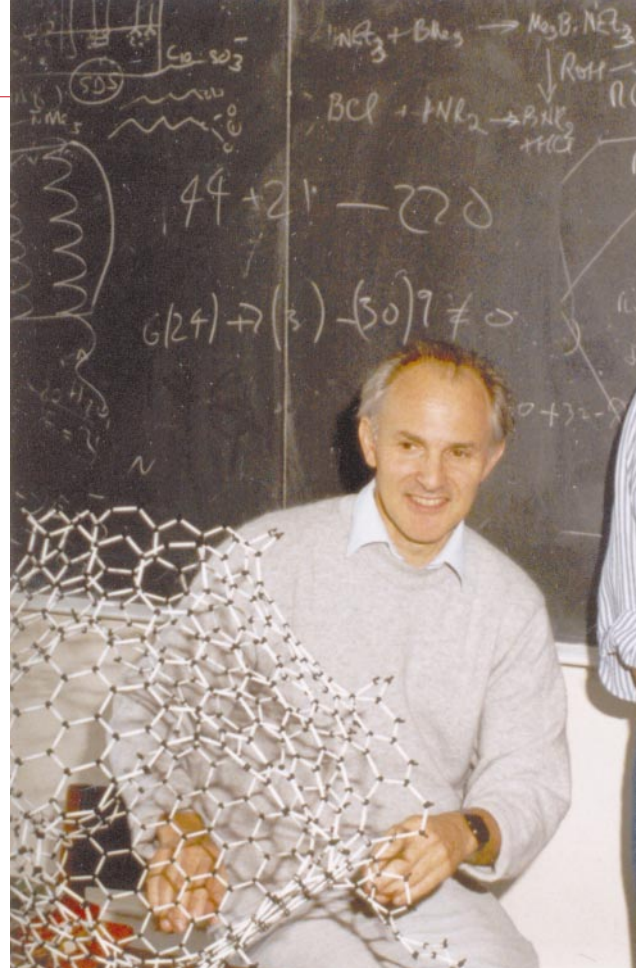
Nanoscience is fragmenting into tinier pieces, but there are great expectations everywhere. Myrna Watanabe investigates.

Nanotechnology is one of those rapidly growing fields that is truly international. Unlike biotechnology, in which the United States took a lead from the start, countries are competing on a more equal basis for a slice of the action in nanotech. For 2004, the US administration is proposing US\$847 million for nanotech funding while Japan and the European Union are not far behind. Taiwan is in the middle of a six-year, US\$620-million national nanotech programme.

Before anyone rushes to switch fields, this needs to be put into perspective. The sums earmarked for nanotechnology research are still a fraction of those spent on biomedical research and development. For example, the US National Institutes of Health's budget for biomedical research this year is \$26.5 billion.

But the nanotech funding is enough to fuel more research, more jobs in academia and discoveries that will spin out into start-up companies or be licensed to existing firms. It is a lot of money for a young, relatively unproven field. And the geographical spread of funding means a broad and varied international job market. Jobs are spread out over a wide area, and are divided up between major academic centres, individuals with small grants, large corporations and hundreds of tiny start-up companies.

The shortage of trained people — caused in part by the field's expanding boundaries — means that even a master's degree is desirable. Microsystems engineer Rajsekhar Popuri was recently hired by the Center for



Ocean Technology of the University of Southern Florida (USF) in Tampa, straight from his master's degree programme at USF. He is continuing his PhD studies part-time.

And a multidisciplinary background certainly helps. Manisha Rane's background includes a PhD in magnetic materials for magnetic recording applications, a guest-scientist stint in Germany on slag corrosion of ceramics in coal-gasification atmospheres, research on high-temperature thermochemistry and a job at the General Electric Global Research Center in Niskayuna, New York. She is now a postdoc at Albany NanoTech, the nanotechnology centre of the State University of New York at Albany.

Stepping into the wider world

Despite the international market in nanotech jobs, some researchers suspect that PhDs from the United States won't be taking up their share of them. "I think that domestic students, although they're interested in the chance to work in labs in different countries, probably take advantage of those opportunities far less than students from Japan, India and Europe," says Evelyn Hu, scientific director of the California Nanosystems Institute on the campus of University of California, Santa Barbara (UCSB). But there will be more and more nanotech jobs outside the United States. And now there is help in an area where US scientists often fall behind their international colleagues — foreign-language skills.

Brian Gerardot, who is completing a PhD on self-assembled quantum dots (crystals several atoms in diameter) under Pierre Petroff and Dirk Bouwmeester at UCSB, will soon move to the Center for Nanoscience at Ludwig-Maximilians-University in Munich, Germany. He doesn't speak German, but has applied to the National Science Foundation to fund six months of language study.

M.W.

INTERNATIONAL OPPORTUNITIES

Because of this shortage, countries are already trying to recruit scientists from abroad. Yury Gogotsi, associate dean of material science and engineering at Drexel University in Philadelphia, says that Japan is trying to recruit foreign postdocs. "I've seen similar ads from Taiwan and Hong Kong, where people anticipate hiring a large number of postdoctoral students," says Gogotsi, who started out in the Ukraine.

There could indeed be more nanotechnology jobs outside the United States than within. But will young US scientists leave to fill them? The trend in other disciplines is for relatively few US scientists to move abroad. But nanotech could be different (see 'Stepping into the wider world', left).

Harry Kroto of the University of Sussex, UK, who shared the Nobel Prize in Chemistry in 1996 for the discovery of buckminsterfullerene, says that

Prizewinner: Harry Kroto's (left) discovery of fullerene structure has opened up a whole new field of carbon nanochemistry.

spin out nanotechnology-based companies. It has already launched several, including two in nanobiotechnology that employ about 100 young scientists and technicians. In Leuven, Belgium, the Interuniversity MicroElectronics Centre is building a micro- and nanoelectronics facility with a E37-million (US\$44-million) grant from the Flemish government and a E46.8 million bank loan. The facility is designed to handle 300-mm silicon wafers for integrated circuit design and will be completed next year.

Singapore has got in on the act with a new nanotech institute housed in the country's 18.5-hectare Biopolis, which opened last month (see *Nature* **425**, 746–747; 2003). The current 90 staff will eventually grow to 250, who will include scientists, technicians, postdocs and graduate students.

The Singapore centre exemplifies the field's fragmentation, as it is focusing primarily on biotech applications for nanotech, such as tiny diagnostics. The country already has a more traditional materials-science centre, which also has a nanotech component.

France is also going into nanobiotech. Minatoc — a e400-million micro- and nanotechnology centre in Grenoble, is hosting Nanobio, a collaboration between the nuclear research agency (CEA) and Joseph Fourier University. Nanobio aims to develop miniaturized

systems for single-molecule tracking, improved drug delivery and analysis.

NEW COMPANIES, NEW JOBS?

Governments funding these big centres want their investment to spin off new companies, and that is starting to happen. Steven Jurvetson, managing director of the venture-capital firm Draper Fisher Jurvetson in Redwood City, California, notes that all 20 nanotechnology-related companies in which his firm has invested came out of university and national laboratories.

Crossing continents: Yuri Gogotsi (left) came west to Philadelphia from the Ukraine, while Brian Gerardot (above) is set to leave California for Europe.

But will those spin-offs create many jobs? The jury is still out. Some point out that few jobs have yet been created worldwide, in part because there are few nanotechnology-related products and the start-ups are so young. One exception may be in the businesses that support the industry, such as specialist lawyers and financiers.

The US National Institute of Standards and Technology estimates that there are about 1,700 nanotechnology companies worldwide. But according to Mark Modzelewski, executive director of the NanoBusiness Alliance in New York: "The overwhelming number in this first phase of nanotech are going to fail beyond belief."

Researchers going into the industry need to weigh the risks and advantages of joining a small start-up. Small companies allow an employee to pick up more skills and perhaps advance more rapidly. Big companies such as IBM are inherently more stable. ■

Myrna Watanabe is a freelance science writer in Patterson, New York.

Web links

US National Nanotechnology Initiative
♦ nano.gov
NSF International Research Fellowship Program
♦ www.nsf.gov/pubs/2002/nsf02149/nsf02149.htm
European Commission's Nanotechnology Website
♦ www.cordis.lu/nanotechnology
California Nanosystems Institute
♦ www.cnsi.ucsb.edu/directory/directory.html
A. J. Drexel Nanotechnology Institute
♦ nano.materials.drexel.edu/DNI
Center for NanoScience
♦ www.nanoscience.uni-muenchen.de

nanotechnologists can come from a variety of disciplines, but must have a specialized skill set. The types of people needed, Kroto says, are those "who really understand how to put atoms together".

CENTRES OF EXCELLENCE

Outside large companies, much of the money for nanotech goes to single investigators, to support a few students and postdocs. These groups are getting specialized. For example, Scot Martin of Harvard University heads a group working on nanogeoscience, which studies environmental processes at minute scales. Some larger centres are focusing on biotech applications, whereas others are more interested in creating exotic materials or tiny microprocessors.

The biggest concentration of academic jobs, however, is in large centres of excellence. In the United States, the National Science Foundation supports eight nanotechnology centres (see *Naturejobs* 4–6, 15 August; 2002). And last month an anonymous donor pledged \$36 million to build a nanotech research centre at the Georgia Institute of Technology in Atlanta, with the state promising a further \$45 million in support over the next few years.

Other large centres around the world are keeping up. The Center for NanoScience at Ludwig-Maximilians-University in Munich, Germany, aims to